Computational Analyses of Immune Responses at Disparate Temporal and Spatial Scales

by

Mikhail Yanislavovich Wolfson

B.S., University of Wisconsin—Madison (2002)

Submitted to the Department of Chemistry
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

September 2012

Author .....
Department of Chemistry
June 14, 2012

Certified by .....
Arup K. Chakraborty, Ph.D.
Robert T. Haslam Professor of Chemical Engineering
Professor of Chemistry and Biological Engineering
Thesis Supervisor

Accepted by .....
Robert W. Field, Ph.D.
Haslam and Dewey Professor of Chemistry
Chairman, Department Committee on Graduate Students

This thesis has been examined by a Committee of the Department of Chemistry as follows:

Thesis Committee Chair .....
Troy Van Voorhis, Ph.D.
Associate Professor of Chemistry

Thesis Supervisor .....
Arup K. Chakraborty, Ph.D.
Robert T. Haslam Professor of Chemical Engineering
Professor of Chemistry and Biological Engineering

Thesis Committee Member .....
Mehran Kardar, Ph.D.
Francis Friedman Professor of Physics

Computational Analyses of Immune Responses at Disparate Temporal and Spatial Scales

by

Mikhail Yanislavovich Wolfson

Submitted to the Department of Chemistry
on June 14, 2012, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

Abstract

In order to perform reliably and protect against unpredictable attackers, immune systems are organized via complex, hierarchical cooperativity. This organization is necessary for their function and a tremendous challenge to their understanding that has motivated contributions from many outside fields. Our approach to studying the immune system computationally has been pragmatic: we have applied any analysis method necessary to understand questions motivated by experimental biology, rather than use biology specifically to discover new physics or methods. Our approach has led us to study problems that span a wide range of time and length scales and require a diverse set of solutions. This thesis describes three projects that span the extremes of this range, from nanometers over nanoseconds to organism-wide responses over hours.

The first project was motivated by a puzzle: experimentalists had reached opposing conclusions on the role of a peptide fragment in the main protein interaction responsible for immune recognition. We used molecular dynamics simulations of the proteins to resolve the contradiction by creating a unifying model.

The second and third projects jump from the molecular to the system-wide. In the second project, we sought to understand which phenotypes of cancer-fighting immune cells were the most important. To do this, we developed novel data visualizations and applied multivariate dimensionality reduction and regression to understand high-dimensional immunotyping data collected on the phenotypes.

The final project addressed an important question in immunology: how accurately do blood assay results reflect the immune response in the tissues, where it matters most? We explored this relationship using a supervised learning model of a highly multidimensional dataset that combined blood and tissue measurements. We found that the two environments could be drastically different and that the relationship mapping blood to tissue was complex.

Combined, these three projects highlight the variety of scientific questions and richness of insight that occur at the intersection of immunology and computation.

Thesis Supervisor: Arup K. Chakraborty, Ph.D.
Title: Robert T. Haslam Professor of Chemical Engineering
Professor of Chemistry and Biological Engineering

Acknowledgments

As much thought as scientists devote to the rational nature of the scientific method, science is a deeply human process. At the core of that process are the interactions between colleagues and mentors, friends and loved ones that pave the way for insight, growth, and learning. To this end, I am grateful to all the people in my life who have served in these capacities for all their help during my education and professional development. As this list is long and my memory finite, I may fail to mention some important members by name, for which I humbly apologize.

I am thankful to my thesis adviser, Professor Arup Chakraborty, for his brilliance, patience, honesty, and generosity. Seeing Arup interact as a scientist is an inspiring and intimidating pleasure. I know of no one else who grasps the essence of an idea so quickly, attacks misconceptions so fiercely, and can rapidly express subtle and correct scientific thinking so eloquently. Learning from the way Arup conducts science and communicates his thoughts has been a continuous inspiration. It has impacted my intellectual growth profoundly. I am also thankful to Arup for his patience and generosity. Every year, he invited anyone in the lab who did not have plans for Thanksgiving to his house. Although I always had other plans, I saw this open invitation as emblematic of Arup's commitment to his students. Arup expressed this commitment to me by patiently allowing me wide latitude to explore and learn and by giving me the opportunity to grow independently, allowing me to take several internships that formed my professional development, even if they were not in academia. For this combination of intelligence and generosity, I am deeply grateful.

In addition to my adviser, I would like to thank the other members of my thesis committee, Professors Mehran Kardar and Troy Van Voorhis. Both Mehran and Troy have taught unassailably brilliant classes. I will always remember the challenge and satisfaction of learning from them. In addition to being amazing teachers, both have been very supportive of my work and provided useful insight in scientific discussion. Troy also served as my committee chair, and through his annual meetings with me was an essential source of wisdom and support at decisive moments in my career.

I would also like to thank all of my scientific collaborators. Kwangho Nam made essential contributions to the molecular dynamics work described in Chapter 2. Without his patience, wisdom, and experience, the simulations I set up may never have succeeded. It was an honor to work with him. It was wonderful to work with Begoña Comin-Anduix and Antoni Ribas to understand the multidimensional FACS data described in Chapter 3. Their careful experiments, enthusiasm for the work and open minds made for a fruitful collaboration. Finally, I owe a great debt of gratitude to Ofir Goldberger who collected the cytokine data analyzed in Chapter 4. Not only is he an amazing scientist, but it has been nothing but a pleasure to work with him. Our conversations have been a free-ranging, collaborative exchange of ideas, and I have deep appreciation for his willingness learn to understand my analyses and to see their potential. I could not have imagined a better collaborator.

In addition to my direct collaborators, my work benefited tremendously from the advice of other professors and scientists. Professor Martin Karplus, apart from his groundbreaking contributions to biophysics, provided thoughtful insight in guiding the molecular dynamics work. And I owe a particular debt of gratitude to Professor Herman Eisen, whose insight into TCR/pMHC binding was of tremendous importance to the work in Chapter 2, and whose amazing depth of vision is paired with a staggering amount of humility and kindness. I have been beyond fortunate to engage with him as a scientist.

The members of the Chakraborty lab have been amazing colleagues. I have shared so much of my life and work with them, and I have been lucky to receive their insights, lean on them for support, and grow intellectually together with them. While every member of the lab has had a positive impact on me, I would like to highlight Abhishek Jha, whose conversations with me provided much needed scientific and emotional support early in my career, and Steve Abel, whose impact on the culture of the lab has been profound. Steve's supportive nature and unabashed love of science has reminded me of why I was inspired by this discipline to begin with. His presence has improved both the productivity and happiness of everyone in the lab.

One of my roles in the lab was to manage our supercomputing cluster. I was aided tremendously in this task by Greg Shomo, who, apart from being a terrific problem solver, was also a pure joy to work with. He was responsive, skilled, and helpful, and he infused the process of computer administration with wit and humor.

Interacting with my classmates in the Department of Chemistry was one of the best parts of the graduate experience. Hiking trips to the White Mountains with Yogesh Surandranath, Peter Allen, Brian Walker, Brandi Cossairt, Jared Silvia, Lee-Ping Wang, Hee-Sun Han, and Kit Werley introduced me and my wife to the beauty of New England. And our international adventure to visit South India for Krupa Ramasesha's wedding with Harold Hwang, Nate Silver, Barratt Park, and Tony Colombo was an unforgettable experience. Sharing our joys and the challenges of coursework and research formed strong bonds among us, and I am proud to consider such intelligent and kind people among my friends.

I owe a great debt of gratitude to my close personal friends outside the department. Wayne Staats, my roommate for 5 years in Wisconsin and Boston, has been my best man in many ways. I am indebted to him for his humor, for his unwavering loyalty, for his love of learning and debate, for his kind support, for his sense of fun, and for being a terrific role model as an engineer, friend, husband, and father. He and his wife Brooke have been close companions to my wife and me, as a fellow young family exploring Boston. Michael Slootsky and Mark Schneider, my other groomsmen, have been invaluable sources of support and friendship, despite living hundreds of miles away. That we have been able to stay close and continue to see one another, or stay in touch with technology despite the distance has brought me deep joy. I have also had the distinct joy of interacting with many other groups of friends. The MechWarriors: John Roberts, Shawn Chester, Dave Craig, Allison Beese, Kevin Cedrone, Andrew Shroll, and Jenna McKown: all MIT mechanical en-

gineers whom I met through Wayne, have been great friends and wonderful companions at trivia competitions and get-togethers. The Friends of the BSO, led by Kartik Venkatram, were wonderful companions for me and my wife through the thriving musical culture of Boston. Our friends Joel and Rachel Adler have been a great source of joy and companionship as another young couple transplanted from Wisconsin to Boston. And we will remember with overwhelming fondness our trips to Mexico to visit our friends Israel Garavito and Lorea Herrera. Their hospitality and generosity made a foreign country feel like a second home.

Over the years leading up to graduate school, I have had many teachers who inspired me and influenced my development as a scientist. At the University of Wisconsin, Professor Qiang Cui was an ideal research adviser and teacher. He treated education very seriously and devoted large amounts of time to my project's development, even though I was only an undergraduate in his lab. I will always appreciate the attention he paid me and be inspired by his broad, precise knowledge and quick thinking. Professor F. Fleming Crim, who taught my freshman chemistry class at the University of Wisconsin, was an overwhelming factor in my decision to pursue physical chemistry as a major and later as a graduate career. His lectures were stunning examples of pedagogy that combined humor, scalpel-sharp clarity and child-like wonder. After I was no longer in his class, he never ceased to provide me with advice and encouragement. And if I trace the path down further, I must mention my fourth grade teacher, John Gildseth. In addition to having a remarkable way of relating to and inspiring his students, John had a gift for seeing the unique potential his students held. When all the other kids wanted to go outside to play at recess, he let me stay inside and play with the Apple IIe in the classroom. Later, he helped our family get a free computer that the school no longer needed. His decision to encourage my interest had a life-changing impact on me. There is no doubt in my mind that his decisions started my trajectory toward the intersection of computers and science.

I'd like to thank the staff of the Chemistry Education Office, in particular, Susan Brighton and Melinda Cerny. Susan served as an indefatigable font of wisdom and support for me and many other graduate students in the department. She always had time to listen, provide good advice, and never pushed. Melinda Cerny took an active role in the students of the department. When I started a student group for thesis writing, she was boundlessly helpful and supportive, providing insight, and helping the idea grow.

In addition to my friends, colleagues, and mentors I want to thank the families who have supported me in addition to my own: my mother- and father-in-law, Kristine and Tom Wendlandt, have welcomed me with open arms, cheered me on and supported me as if I were one of their own. Margaret and Al Timmerman, lifelong family friends, have played a tremendous role in my development. In addition to supporting our family after we immigrated to the United States, they continue to serve as inspiration and guidance. I treasure every opportunity I have to see and speak with them.

It is traditional to thank one's immediate family last. I can understand this tradition: perhaps its originators, like me, found it difficult to adequately give

thanks for such a large debt of gratitude and so could only address it at the end, when they were forced to confront it. Although I know anything I write here will be vastly inadequate, I will try to convey my gratitude below.

I owe an immeasurable amount of gratitude to my wife Johanna Wolfson. In may ways, the story of graduate school has been a story about us. We came to MIT together and entered the same program. We took and taught many of the same classes, prepared for the same exams, and stayed awake the same sleepless nights learning, working, preparing together. All of our shared experiences brought us even closer, and in the middle of graduate school, we got married. Throughout this challenging and inspiring process, I could not have been blessed to with a more loving, supportive, or brilliant partner. Jo has shared in my triumphs, supported me through my setbacks and saved the day on countless occasions with her quick wit, sound advice, kind words, and brilliant plans. She brings out the best in me by always striving for the best, and she has brought beauty, humor, friendship, and love into what could otherwise be an isolating process. Her own emergence as an inspiring leader, her amazing intellect, and her sense of duty and fairness have formed an indelible impression on me and continue to inspire me to grow as a person. Thank you, Jo, for every single day we've been able to spend together.

I am also deeply thankful to my family of origin, my parents, Yanislav and Inna Wolfson, and my brother Boris Wolfson. I want to acknowledge my parents in particular. I was born in the city of Yalta, part of the Soviet Union, in 1984. My parents sacrificed tremendously, leaving their parents, friends, and careers behind to bring our family to the United States in 1990. Through their sacrifice, they created unparallelled opportunity for our family. Through their undaunted work ethic, unbridled love of learning, and respect for truth, they transformed the opportunity into a successful life and inspired my brother and me to never cease learning and work hard for success. Not only did they raise me: they intentionally decided to change everything in my life for the better, at great cost to themselves, so that instead of growing up oppressed, I might grow up free. The debt I owe them is impossible to describe, much less repay.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

This thesis summarizes several years of work toward understanding the immune system with computational tools and simulations. The immune system is one of the most complex areas of biology imaginable. Because it deals directly with fighting disease, and its aberrant behavior is responsible for so many other diseases, it is also at the forefront of a great amount of medical research. Advances in immunology have tremendous potential to relieve the untold amount of suffering caused by viruses such as HIV and autoimmune disorders, such as diabetes. With the emergence of computational biology and biophysics, engineers, chemists, physicists and computer scientists have all been enticed by immunology. Our work falls into this greater context. Apart from the potential human benefits and the inherently interesting biology, immunology offers quantitative researchers a tremendous field of potential and complexity to parse. The system works through hierarchically structured cooperativity, which means that significant processes happen in a variety of temporal and spatial scales. All of these processes must be made clearer, with new techniques and new analyses, and it is this rich world of emergent behavior and nontrivial interaction that draws quantitative scientists in.

1.1 The immune system

The immune system is essential to higher organisms' survival. It protects its host from invading viruses and other pathogens through a wide ensemble of defense methods. The methods can be divided into two major parts: the innate immune system, which is built to recognize common patterns of pathogens and destroy them, and the adaptive immune system, which trains a diverse repertoire of cells (T lymphocytes) to recognize pathogens that may never have been seen before, and remember any new pathogens encountered.

Our work is concerned mostly with the adaptive immune response. In this response, the pathogen is absorbed by the host's own cells, either via infection or phagocytosis by innate immune cells. The pathogen's proteins are cut up into short peptide fragments and displayed as antigen on the surface of the consuming cell, held in bun-like structures of surface receptors called MHCs. When T lymphocytes encounter the surface of the antigen-presenting cells, their specialized receptor (TCR), binds to the peptide-MHC complex [1,2]. If the reaction is sufficiently strong, the T cell detects the presence of infection and begins to orchestrate a series of effector functions [3].

After detection occurs, the immune system organizes a complex, multi-pronged response, incorporating many cell types and many strategies for pathogen control. The amounts of coordination necessary to implement such a response are achieved through the help of cytokines: small inter-cellular proteins released as signals. Common examples are IL-2, a cytokine which signals T cell proliferation, and TGF-β, a cytokine that usually inhibits an inflammatory response. Cytokines exist in a wide variety of forms and send a wide variety of messages: the same cytokine in different combinations can have different effects [4]. Through their variety and complexity, cytokines provide a rich window into the functions of the innate and adaptive immune system. Their study is of particular interest to a sys-

tematic understanding of immune responses at larger scales.

1.2 Scope of the work and choice of methods

Computational immunology is a tremendously varied field, emerging from the need to systematically understand the tiered complexity of the immune system and to interpret the ever more complicated laboratory results emerging today. Recent efforts in the field have combined advanced experimental techniques with analysis techniques originally from financial analysis, stochastic processes, chemometrics, machine learning, chemical kinetics, and neurobiology [4–12].

The scope of our work does not lie in developing new contributions in any of the original fields from which these techniques came, it does not lie in developing new techniques for their own sake, and it does not lie in immunology experiments. Rather, our contributions come from bringing computational techniques to bear on relevant immunological questions, often in novel ways. The question we ask is, "how can we use currently existing techniques to answer the *biologically relevant question*?" Such synthesis of computational techniques with immunological data, ever with an eye toward biological relevance, can produce novel insight and bring new interpretability to the imaging and multiplexing techniques that have recently revolutionized the field.

The optimal method to use in solving a problem depends on the questions being asked. For questions at the molecular level that involve individual proteins interacting with one another, molecular dynamics is a highly appealing solution for its ability to provide a direct answer. For questions that deal with the immune system as a whole, a variety of methods exist. Within the large scope of systematic immune questions, one frequently recurring problem is to find a way to model, explore, and understand highly multidimensional data, which does not lend it-

self readily to traditional analysis methods or charts. Multivariate statistics and machine learning offer a powerful tool set for these tasks.

Our work seeks to answer two classes of problems. The problems differ greatly from one another and serve as bookends for the kinds of temporal and spatial scales addressed in contemporary computational immunology. The first type of problem involves the impact of a peptide on protein binding. Its spatial scale is nanometers and its timescale is nanoseconds, making molecular dynamics simulations an appropriate technique to use. The second class of problem requires the analysis of multidimensional data collected at multiple timepoints throughout a human or mouse immune system. Its spatial scale spans the entire organism, and its time scale can extend to days. Since the goal is to understand and interpret data that has already been collected, multivariate statistical techniques are the right class of methods to deal with this class of problem.

The remainder of this chapter explores the techniques outlined above in greater detail. Chapter 2, which is adapted from a paper published in the *Journal of Physical Chemistry B* [9], describes our use of molecular dynamics to understand the subtle influence of the peptide on the interface between TCR and allo-MHC, covering the first class of problem. Chapters 3 and 4 conclude the thesis and describe the other end of the spectrum of our work, the second class of problem. They are pre-print manuscripts of two detailed case studies of the successful application of dimensionality reduction and latent variable regression analyses to highly multidimensional immunological datasets.

1.3 Molecular dynamics simulation

When the answer to an immunological question rests on the interactions between proteins, molecular dynamics provides a way to answer it. The technique was orig-

inally developed in pure physical simulation [13]. Since its inception, however, it has been used heavily in more applied fields, such as material science [14] and biology [15–17]. Molecular dynamics attempts to describe the structural and dynamical properties of a system by solving its classical equations of motion numerically, given a starting state. In biology, the system in question is a large biomolecule and potentially its surrounding environment. Every atom in the system is described classically. The initial position is usually obtained from a crystal structure, or a homology or docking model based on another crystal structure. With molecular dynamics, we can answer the question, "How do these proteins move?" We can also sometimes answer the question, "If we change this part of a system, what will happen?" This lets us address many questions at the heart of immunology for any protein with a sufficiently accurate crystal structure, such as many types of MHCs and TCRs [18, 19], signaling molecules such as Ras [20], or cytokines [21].

The chaotic nature of the classical equations of motion, combined with the large number of atoms to resolve mean that with current computing power, molecular dynamics can typically address scales on the order of hundreds of monomers, for on the order of tens of nanoseconds [22]. Extension of these scales by more than an order of magnitude is possible with a variety of coarse graining and approximation techniques, which are beyond the scope of this introduction.

When molecular dynamics simulations are performed on biomolecules, the molecules are modeled as systems of atoms with mass and charge, connected by spring-like bonds and subject to a host of other long and short-range interactions, such as dihedral angles, electrostatic interactions, and Van Der Waals interactions. These interactions are parameterized for each atom type considered and tabulated in a comprehensive data structure called the force field. Different force fields contain different parameter values, use different types of terms, assume different molecule topologies. As a concrete example, we will use the CHARMM poten-

tial [23], a widely used force field. In the most direct form of dynamics with the CHARMM potential, each atom i with position $\boldsymbol{r}_i$ is subject to the classical equation of motion

$$m\ddot{\boldsymbol{r}}_i = \boldsymbol{F}_i = -\boldsymbol{\nabla}_i V(\boldsymbol{r}_1, \ldots, \boldsymbol{r}_N), \tag{1.1}$$

where V is determined by the CHARMM potential:

$$\begin{aligned} V(\boldsymbol{r}_1, \ldots, \boldsymbol{r}_N) = V(\boldsymbol{R}) = \underbrace{V_\mathrm{b}(\boldsymbol{R}) + V_\theta(\boldsymbol{R}) + V_\varphi(\boldsymbol{R}) + V_\omega(\boldsymbol{R})}_{\text{bond terms}} \\ + \underbrace{V_\mathrm{el}(\boldsymbol{R}) + V_\mathrm{VdW}(\boldsymbol{R})}_{\text{non-bond terms}} + \underbrace{V_\mathrm{restr}(\boldsymbol{R})}_{\text{restraints}} \end{aligned} \tag{1.2}$$

The individual terms represent the different types of bonding and non-bonding interactions the atoms in the system are subject to, plus any external restraints. We address the terms individually, treating the tabulated parameters of the force field as functions of atom index of the form $f(i, j, k, l)$.

The bonding term V_b is a simple harmonic spring:

$$V_\mathrm{b}(\boldsymbol{R}) = \sum_{(i,j)\in\mathbf{B}} k_b(i,j)\Big[r_{ij} - r_0(i,j)\Big]^2, \quad r_{ij} = |\boldsymbol{r}_i - \boldsymbol{r}_j|, \tag{1.3}$$

where $\mathbf{B}$ is the set of all bonded atom pairs in the structure, and k_b and r_0 are force constants and equilibrium bond lengths, parameterized in the force field by the types of atoms i and j.

The angle term V_θ is also a harmonic spring, but now in the angle space:

$$V_\theta(\boldsymbol{R}) = \sum_{(i,j,k)\in\mathbf{\Theta}} k_\theta(i,j,k)\Big[\theta(\boldsymbol{r}_i, \boldsymbol{r}_j, \boldsymbol{r}_k) - \theta_0(i,j,k)\Big]^2, \tag{1.4}$$

where $\mathbf{\Theta}$ is the set of all bond-bond angles in the structure, $\theta(\boldsymbol{r}_i, \boldsymbol{r}_j, \boldsymbol{r}_k)$ is the angle between the three atoms, and k_θ and θ_0 are force constants and equilibrium bond angles, parameterized as above.

The proper dihedral angle term V_φ represents the gauche steric effects of atoms rotating around a bond with a periodic term:

$$V_\varphi(\boldsymbol{R}) = \sum_{(i,j,k,l)\in\mathbf{\Phi}} |k_\varphi(i,j,k,l)| - k_\varphi(i,j,k,l)\cos\Big[n(i,j,k,l)\varphi(\boldsymbol{r}_i, \boldsymbol{r}_j, \boldsymbol{r}_k, \boldsymbol{r}_l)\Big], \tag{1.5}$$

where $\mathbf{\Phi}$ is the set of all dihedral angles in the structure, $\varphi(\boldsymbol{r}_i, \boldsymbol{r}_j, \boldsymbol{r}_k, \boldsymbol{r}_l)$ is the dihedral angle including the four atoms, k_φ is the force constant, and $n \in \{1, \ldots, 6\}$ specifies how many groups are rotating around the bond.

The improper dihedral angle term V_ω is a harmonic restraint in dihedral space added to maintain chirality and planarity in certain configurations, such as carbonyl groups [15]:

$$V_\omega(\boldsymbol{R}) = \sum_{(i,j,k,l)\in\mathbf{\Omega}} k_\omega(i,j,k,l)\Big[\omega(\boldsymbol{r}_i, \boldsymbol{r}_j, \boldsymbol{r}_k, \boldsymbol{r}_l) - \omega_0(i,j,k,l)\Big]^2, \tag{1.6}$$

where $\mathbf{\Omega}$ is the set of all improper dihedrals in the structure, $\omega(\boldsymbol{r}_i, \boldsymbol{r}_j, \boldsymbol{r}_k, \boldsymbol{r}_l)$ is the improper dihedral including the four atoms, and k_ω and ω_0 are force constants and equilibrium bond angles, parameterized as above.

The non-bond terms are calculated pairwise for the atoms in the system, which grows as N^2. These terms are often the most computationally expensive to calculate and are subject to simplification with cutoffs and scaling of shifting algorithms that are also beyond the scope of this introduction, but are explained in detail in the pertinent papers [15, 23]. In the following equations, these algorithms will be represented by the scale factor $\delta_{ij} \geq 0$, which depends on the algorithm. For atoms that do not interact, $\delta_{ij} = 0$.

The electrostatic non-bond terms in the CHARMM potential are calculated with Coulomb's law:

$$V_{\text{el}}(\boldsymbol{R}) = \sum_{i=1}^{N} \sum_{j=1}^{N} \delta_{ij} \frac{q(i)q(j)}{4\pi\epsilon r_{ij}}, \tag{1.7}$$

where q represents atoms' charge, r_{ij}, as defined earlier, is the distance between them, and ϵ is the dielectric constant, which may be changed depending on the environment, or given functional dependence.

The Van der Waals interactions responsible for the London dispersion forces between the atoms are approximated by the Lennard-Jones potential [24]:

$$V_{\text{VdW}}(\boldsymbol{R}) = \sum_{i=1}^{N} \sum_{j=1}^{N} \delta_{ij} \left[\frac{A(i,j)}{{r_{ij}}^{12}} - \frac{B(i,j)}{{r_{ij}}^{6}} \right], \tag{1.8}$$

where A and B are parameterized values that are simple functions of the dielectric and the Van der Waals radii of atoms i and j.

Once the initial position is set and the parameters are known, initial velocities are assigned to the atoms, most often from a Maxwell-Boltzmann distribution, solvent is added, if necessary, and the equations of motion are solved as a trajectory. The positions of the atoms and any other properties of interest are periodically recorded for later analysis. After the trajectory has equilibrated, we can run it further to observe realistic behavior of the protein by viewing movies of the dynamics. From the equilibrated section of the trajectory, we can also calculate structural properties, correlation functions, and thermodynamic quantities to obtain a complete physical description of the system.

Because of the starkness of the approximations and assumptions, molecular dynamics calculations do not always provide accurate results. Sometimes, this is due to a poor initial condition from an insufficiently defined model of the structure. Sometimes, this is due to a bad choice of parameters. Despite the many potential opportunities for error, simulations can be verified by comparing to known results

and thereby serve as reliable indicators of molecular behavior.

For all of the above-mentioned strengths, molecular dynamics simulations have been used many times to explain the behavior of key immunological proteins [25–30] and were the ideal calculations to perform to understand the role of the peptide in alloreactivity.

1.4 Dimensionality reduction and latent variable regression methods

As explained above, Molecular Dynamics can serve as an incredibly effective window into the behavior of the key proteins involved in immunology. But neither it—nor any other type of simulation—is a complete solution. Often, there is not enough data to permit the construction of a model that can be simulated. More importantly, the questions asked may not be appropriate for simulation.

Another type of question that frequently emerges in immunology is to find the overall structure and most important variables in a highly multidimensional dataset. Answering this question requires techniques such as exploratory data analysis (unsupervised learning) [31] and regression (supervised learning), often in combination. When successful, these techniques identify which relationships and degrees of freedom are the most important, which observations are outliers, and sometimes construct a consistent model of the data that elucidates the underlying biology and suggests further experiments.

The recent need for regression and exploratory analysis is spurred by the concomitant rise in multiplexing instruments, which can simultaneously measure many analytes [32, 33], and systematic experimental approaches, which seek to understand a biological system by measuring a diverse set of indicators [34–37]. However, no advance in instrumental technology or approach can alleviate one of

the fundamental difficulties of biological experiments: data from living organisms. No matter how much information is collected about a particular mouse, an experiment can only use a small number of them. The live organism constraint means that, in addition to many dimensions, the new datasets have few observations. Such datasets are a challenge to analyze, because of the difficulty of understanding multidimensional variation and because of the large potential in variability that comes from small sample sizes. Fortunately, robust statistical methods have been developed for regression and exploratory analysis of these datasets. Our work has aimed to bring clarity to complex immunological phenomena through judicious application of these simple yet powerful techniques.

The term *exploratory analysis* is a broad umbrella that encompasses many techniques, including a multitude of clustering approaches. Within that umbrella, we have decided to focus on latent variable techniques for dimensionality reduction. These approaches seek to find new (latent) variables, that capture the essential aspects of the data in fewer dimensions. The data, once projected into the lower-dimensional representation, can then be subjected to regression techniques to identify key relationships among the variables. The simplicity of the techniques comes from the straightforward, linear nature of both the latent variable calculation and the regression. It may seem counterintuitive at first to model potentially nonlinear data with such a strictly linear model. But the benefits that linearity brings to interpretation are unparalleled. And given the qualitative nature of exploratory analysis, a linear analysis is often sufficient: what matters is not the most accurate model, but an understanding of the essential variables and degrees of freedom, a qualitative picture of a complex system. For this purpose, the techniques we highlight below are ideal.

1.4.1 Principal component analysis

Principal component analysis (PCA) is a basic exploratory data analysis technique. It is one of the simplest ways to project a dataset onto a useful set of reduced dimensions. Because of its simplicity, is ubiquitous in many fields [38], and its key ideas underlie much of the rest of the techniques used in our work.

The essential idea behind PCA is, given data in many dimensions, to create a new, smaller set of latent variables that span as much of the data as possible. These latent variables are called principal components (PCs). Looking at the data in the space of the PCs can reveal interesting grouping and clustering of the data. Identifying which of the original (manifest) variables contribute to the PCs can reveal correlated groupings of the manifest variables that display the most impact in the data. The PCs are defined as linear combinations of the manifest variables. In order to span as much of the data as possible, the latent variables are determined by an optimization process subject to two essential constraints:

1. every PC must have unit norm, *i.e.* the squares of the weights must sum to 1.

2. all of the PCs must be completely perpendicular to one another

There are many potential criteria to optimize subject to these constraints, but PCA chooses to optimize variance. With the optimization criterion and constraints defined, the PCs are calculated one-by-one. That is, the first principal component of a dataset is a unit-norm linear combination of the manifest variables with maximum variance. The second principal component is the unit-norm linear combination of the manifest variables with maximum variance that is orthogonal to the first PC. The third PC is the same, orthogonal to both the first and second PC, and so on. In principle, there are as many PCs as there are original variables. But the amount of variance remaining for each consecutive PC shrinks drastically, so the data is usually projected onto only the first few.

The discussion is much clearer in mathematical terms. In our analysis, we will assume that each manifest variable in the data is already mean-centered. This simplifies the mathematics without restricting the generality of the technique. For a completely rigorous mathematical treatment, we recommend the work of Jolliffe [38]. Thus, given an $n \times m$ matrix of mean-centered data $\boldsymbol{X}$, whose columns are the m manifest variables $\left(\boldsymbol{X} = \left[\boldsymbol{x}^{(1)} \dots \boldsymbol{x}^{(m)}\right]\right)$, and whose rows correspond to individual observations or samples, PCA seeks to find the matrix $\boldsymbol{W} \in \mathbb{R}^m \times \mathbb{R}^a$, $a \leq m$, such that $\boldsymbol{T} = \boldsymbol{X}\boldsymbol{W}$ is the data projected into the reduced ($a \leq m$) dimensional representation. The $n \times a$ matrix $\boldsymbol{T}$ is known as the score matrix, and $\boldsymbol{W}$ is known as the loading matrix. The orthogonality and normalization constraints are summarized in the relationship $\boldsymbol{W}^T\boldsymbol{W} = \boldsymbol{I}_a$, where $\boldsymbol{I}_a$ is the $a \times a$ identity matrix.

A straightforward derivation [38] shows that $\boldsymbol{W}$ can be obtained by calculating the eigenvectors of the covariance matrix $\boldsymbol{K} = \frac{1}{n-1}\boldsymbol{X}^T\boldsymbol{X}$, such that $\boldsymbol{W}_0^T\boldsymbol{L}\boldsymbol{W}_0 = \boldsymbol{X}$, where $\boldsymbol{W}_0$ is the matrix of eigenvectors, and $\boldsymbol{L}$ is a matrix with only the corresponding eigenvalues on the diagonal. Intuitively, this eigenvalue decomposition makes sense. $\boldsymbol{K}$ represents all of the variance and covariance in the data. Its diagonal elements are the variances of each individual variable. Because the columns of $\boldsymbol{X}$ are mean-centered,

$$\operatorname{var}\left(\boldsymbol{x}^{(i)}\right) = \frac{1}{n-1}\sum_{\ell=1}^{n}\left(x_\ell^{(i)}\right)^2 = \frac{\boldsymbol{x}^{(i)} \cdot \boldsymbol{x}^{(i)}}{n-1} = (\boldsymbol{K})_{ii} = k_{ii}.$$

Its off-diagonal elements are, by analogy, the covariances

$$\operatorname{cov}\left(\boldsymbol{x}^{(i)}, \boldsymbol{x}^{(j)}\right) = \frac{1}{n-1}\sum_{\ell=1}^{n} x_\ell^{(i)} x_\ell^{(j)} = \frac{\boldsymbol{x}^{(i)} \cdot \boldsymbol{x}^{(j)}}{n-1} = (\boldsymbol{K})_{ij} = k_{ij}.$$

Geometrically, the eigenvectors of $\boldsymbol{K}$ represent the set of axes within $\boldsymbol{K}$ with the greatest possible extent, indicated by their eigenvalue. Therefore, just as the eigenvalues with the largest eigenvalues represent the most essential spatial dimensions

of a matrix, the principal components represent the dimensions in the data with the greatest variation in them. Variation is not a perfect proxy for significance, but it is a simple and effective marker nonetheless.

Because $\boldsymbol{K}$ is a square, symmetric matrix, the PCs, i.e., the columns of $\boldsymbol{W}_0$, are defined to be orthonormal. Furthermore, the variance of every PC is equal to the eigenvalue of the corresponding eigenvector, so the first a PCs ($\boldsymbol{W}$) can be obtained by simply sorting the columns of $\boldsymbol{W}_0$ by eigenvalue and keeping the first a columns.

The eigenvalue definition can be extended further with singular value decomposition (SVD). Because the left-singular vectors of an SVD of a matrix $\boldsymbol{M}$ are equivalent to the eigenvectors of the matrix $\boldsymbol{M}^T\boldsymbol{M}$, $\boldsymbol{W}$ can be obtained directly from the left singular vectors of an SVD of $\boldsymbol{X}$. The missing scale factor of $\frac{1}{n-1}$ only affects the eigenvalues and it can be compensated for directly. This approach has the added benefit of automatically putting the most significant vectors first.

If all of the eigenvectors are kept in $\boldsymbol{W}$ ($a = n$), then our original definition of the score matrix $\boldsymbol{T}$ rearranges into a decomposition:

$$\begin{aligned} \boldsymbol{T} &= \boldsymbol{X}\boldsymbol{W} \\ \boldsymbol{T}\boldsymbol{W}^T &= \boldsymbol{X}\boldsymbol{W}\boldsymbol{W}^T \\ \boldsymbol{X} &= \boldsymbol{T}\boldsymbol{W}^T. \end{aligned} \tag{1.9}$$

When not all components are kept ($a < n$), this decomposition is not exact, but it nevertheless serves as a useful reformulation of PCA in terms of a model of the data. The first a PCs are thus the a first and most most significant components in the spectral decomposition of $\boldsymbol{X}$ [38].

The score matrix $\boldsymbol{T}$ represents the data in the lower-dimensional space of the principal components. By exploring this lower-dimensional space visually or with other algorithms, it is possible to detect groups of related observations, or deter-

mine patterns. One common pattern is for a PC to partition the observations into distinct groups. This separation arises naturally, from the correlations in the data, rather than any human involvement or parameterization. Another common pattern is the emergence of clusters with a space of several PCs. These patterns are usually detected visually through the use of scatterplots and biplots [39].

The loading matrix $\boldsymbol{W}$ represents the PCs in terms of the original manifest variables. A column in $\boldsymbol{W}$ shows the loadings that define a PC. The sign convention for each column is arbitrary, so the absolute sign of a loading is unimportant: only differences between signs and magnitude of loadings matter. The first PC represents the overall amount of covariation in the data and in most cases will have universally positive or negative loadings for all of the variables. Often, the columns will highlight certain variables with large loadings. Some will draw contrasts between related groups of variables by assigning them opposite signs. The loadings thus reveal information about the manifest variables in two ways. First, the more relevant a variable is in an early PC, the more likely it is to be associated with some important underlying degree of freedom in the data; second, more than simply identifying the important varibles, the groups of loadings identify variables that vary similarly. This dual purpose makes PCA a powerful analysis technique despite its simplicity.

1.4.2 Multiple linear regression

While PCA is a tremendously powerful technique for finding important correlations among a single block of data, we often want to find relationships between two blocks of data and perhaps to build a model that can capture that relationship. This general problem is called regression, or supervised learning, and for our purposes, we will restrict ourselves to linear relationships.

The linear regression problem seeks to find a matrix $\boldsymbol{B}$ such that $\hat{\boldsymbol{Y}} = \boldsymbol{XB}$ is

optimally close to $\boldsymbol{Y}$ according to some measure, usually a sum ε of squared error terms:

$$\begin{aligned} \boldsymbol{E} &= \left(e_{ij}\right) = \boldsymbol{Y} - \hat{\boldsymbol{Y}} \\ \varepsilon &= \sum_{i=1}^{n} \sum_{j=1}^{m_Y} {e_{ij}}^2. \end{aligned} \tag{1.10}$$

The simplest solution to this problem is to use all the data in $\boldsymbol{X}$ and $\boldsymbol{Y}$ to calculate $\boldsymbol{B}$ directly, since it can be derived that the expression $\boldsymbol{B} = \left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\boldsymbol{Y}$ minimizes ε. However, this approach is frequently inadequate to deal with noisy or co-linear data. In the latter case, $\boldsymbol{X}^T\boldsymbol{X}$ is singular, so $\boldsymbol{B}$ is necessarily undefined. In the former case, because *all* of the data are used in the fit, noisy data are fit as well, which easily leads to over-fitting. As for its practical uses, linear regression is too ubiquitous to mention in depth, due to its simplicity and ease of implementation [40].

1.4.3 Principal component regression

Multiple strategies are in use for resolving the noise and collinearity issues, and a common family of strategies is known as latent variable regression. Latent variable regression methods project $\boldsymbol{X}$ and/or $\boldsymbol{Y}$ onto a (usually smaller) set of latent variables, defined and optimized as above.

The simplest latent variable regression method is called principal component regression (PCR) [38, 41]. Here, the latent variables are principal components (PCS) of $\boldsymbol{X}$, as calculated by PCA [38]. The PCS are calculated from a singular value decomposition (SVD) of $\boldsymbol{X}$, and $\boldsymbol{Y}$ is fit to a subset of the a highest-scoring PCS. Because the PCS are orthogonal to one another, PCR eliminates the problem of co-linear variables. And if the largest directions of variation in $\boldsymbol{X}$ overlap well with the variation in $\boldsymbol{Y}$, PCR can provide robust and accurate models. However, this is rarely the case

in real life: there is nothing that ensures that the principal components of $\boldsymbol{X}$ overlap with variation in $\boldsymbol{Y}$ at all. In fact, the first principal components of real $\boldsymbol{X}$ datasets are frequently devoted to outlier detection, which categorically does not relate to $\boldsymbol{Y}$. So, while PCR can be a simple and effective technique in some situations, it often falls short.

1.4.4 Partial least squares projection onto latent structures (PLS)

Partial least squares projection onto latent structures (PLS) is an effective strategy for overcoming the limitations of PCR at the cost of increased model complexity. Instead of, as in PCR, constructing a reduced-dimensional representation of $\boldsymbol{X}$ from an SVD of $\boldsymbol{X}$, PLS constructs it from an SVD of the cross-correlation matrix $\boldsymbol{K} = \boldsymbol{X}^T\boldsymbol{Y}$ [42]. This approach to finding a PLS regression (SIMPLS) is shown in Algorithm 1.

Unlike PCR, PLS produces two sets of scores and loadings, although both are based on $\boldsymbol{T}$, the set of $\boldsymbol{X}$ scores. The benefit of this more elaborate output is a flexible and powerful algorithm for modeling diverse datasets. The optimal number of components $a^\star$ can be set deterministically through cross-validation. Once this number is determined and the algorithm is run a final time, the model that results contains two decompositions and one regression.

x score-loading decomposition, error term:	$\boldsymbol{X} = \boldsymbol{T}\boldsymbol{P}^T + \boldsymbol{E}$	(1.11)
y score-loading decomposition, error term:	$\boldsymbol{Y} = \boldsymbol{U}\boldsymbol{Q}^T + \boldsymbol{F}$	(1.12)
Predicting $\boldsymbol{Y}$ from the x-scores:	$\boldsymbol{Y} = \boldsymbol{T}\boldsymbol{Q}^T + \boldsymbol{G}$	(1.13)
Predicting $\boldsymbol{Y}$ directly from $\boldsymbol{X}$:	$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{B} + \boldsymbol{G}$	(1.14)

The decompositions above work exactly like the PCA decomposition described earlier, even if the loadings are obtained in a slightly different process.

Algorithm 1 SIMPLS

Require: Data matrices $\boldsymbol{X}$, $\boldsymbol{Y}$, and a maximum dimension number a.
Ensure: The matrices $\boldsymbol{P}, \boldsymbol{Q}, \boldsymbol{T}, \boldsymbol{U}, \boldsymbol{W}$, and $\boldsymbol{B}$ fully specify a PLS model of Y vs X.

$\boldsymbol{K} \leftarrow \boldsymbol{X}^T\boldsymbol{Y}$
for all $i \in \{1, \ldots, a\}$ **do**
  $\boldsymbol{W}\boldsymbol{S}\boldsymbol{C}^T = \boldsymbol{K}$ ▷ Compact SVD of $\boldsymbol{K}$
  $\boldsymbol{w}_i \leftarrow$ the first column of $\boldsymbol{W}$
  $\boldsymbol{t}_i \leftarrow \boldsymbol{X}\boldsymbol{w}_i$ ▷ x-scores for component i
  $\boldsymbol{t}_i \leftarrow \boldsymbol{t}_i / \left(\boldsymbol{t}_i{}^T \boldsymbol{t}_i\right)$ ▷ Normalize $\boldsymbol{t}_i$
  $\boldsymbol{p}_i \leftarrow \boldsymbol{X}^T\boldsymbol{t}_i$ ▷ x-loadings for component i: a least-squares fit of $\boldsymbol{X}$ to $\boldsymbol{t}_i$
  $\boldsymbol{q}_i \leftarrow \boldsymbol{Y}^T\boldsymbol{t}_i$ ▷ y-loadings come from a least-squares fit of x-scores!
  $\boldsymbol{u}_i \leftarrow \boldsymbol{Y}^T\boldsymbol{q}_i$ ▷ y-scores for component i
  $\boldsymbol{V} \leftarrow$ an orthogonal basis of i vectors constructed from $\{\boldsymbol{p}_i\}$.
  $\boldsymbol{K} \leftarrow \boldsymbol{K} - \boldsymbol{V}$
end for
$\boldsymbol{P} \leftarrow [\boldsymbol{p}_1 | \ldots | \boldsymbol{p}_a]$ ▷ Combine into matrices
$\boldsymbol{Q} \leftarrow [\boldsymbol{q}_1 | \ldots | \boldsymbol{q}_a]$
$\boldsymbol{T} \leftarrow [\boldsymbol{t}_1 | \ldots | \boldsymbol{t}_a]$
$\boldsymbol{U} \leftarrow [\boldsymbol{u}_1 | \ldots | \boldsymbol{u}_a]$
$\boldsymbol{W} \leftarrow [\boldsymbol{w}_1 | \ldots | \boldsymbol{w}_a]$
$\boldsymbol{B} \leftarrow \boldsymbol{W}\left(\boldsymbol{T}^T\boldsymbol{T}\right)^{-1}\boldsymbol{Q}^T$ ▷ The regression matrix

Although PLS models incorporate variation from $\boldsymbol{Y}$, the projection is still more fundamentally dependant on the $\boldsymbol{X}$ because the y scores and loadings $\boldsymbol{U}$ and $\boldsymbol{Q}$ come directly from the x scores $\boldsymbol{T}$. Such a restriction does not preclude a diverse world of data, however, and PLS, which was applied first to chemometric analysis of spectra [43] has successfully been used to model everything from human taste preferences [44] to cell signaling networks [34, 35, 37].

Despite their ubiquity, flexibility and power, PLS models can still produce meaningless, uninterpretable results. A key factor necessary for the interpretability of results is score-loading correspondence [45], the similarity of the loadings $\boldsymbol{P}$—obtained from a least-squares fitting of the data $\boldsymbol{X}$ onto the scores $\boldsymbol{T}$—to the original weights $\boldsymbol{W}$ responsible for making the scores. For the model to make sense, $\boldsymbol{W}$ and $\boldsymbol{P}$ should correspond well. In PCA, because of the NIPALS calculation algorithm, $\boldsymbol{P}$ and $\boldsymbol{W}$ are equal by definition [46]. In PLS, however, this is no longer true, because the information from $\boldsymbol{Y}$ impacts the loadings. Because of the correlation, any parts of $\boldsymbol{X}$ uncorrelated with $\boldsymbol{Y}$ act as structured noise, which is still captured by the model. Such noise is problematic. While PLS models are robust to a certain degree of noise, too much noise, particularly structural noise, can make interpretation of the loadings much more difficult [45]. Sometimes, the noise is limited to only certain variables (columns) in $\boldsymbol{X}$ or $\boldsymbol{Y}$, which has led to a host of complex optimization techniques for removing "noisy" variables from the regression data [10, 47–55]. These algorithms add many parameters and long simulation times to the process. Due to the exponentially large number of possible variable combinations, it is usually impossible to be certain of picking a truly optimal variable subset. A simpler, more informative, and more effective approach is instead to filter out this noise directly, which the extensions of PLS O-PLS and O2-PLS do.

1.4.5 O-PLS and O2-PLS

The core of O-PLS, and its sibling O2-PLS, which allows for multidimensional y data, is to use the underlying structure of of PLS, but to improve the interpretability of the model by adding an Orthogonal Signal Correction (OSC) filter to remove the structured noise in both $\boldsymbol{X}$ and $\boldsymbol{Y}$ [45, 56]. The OSC filters deflate the data matrices by anything orthogonal to the correlation between them. Since the structured noise in the data is by definition uncorrelated with the relationship between the data, the filters capture and remove much of the structured noise. This splits the model up into two parts: a predictive part, consisting entirely of the correlations between $\boldsymbol{X}$ and $\boldsymbol{Y}$, and an orthogonal part, consisting of the noise in $\boldsymbol{X}$ and the noise in $\boldsymbol{Y}$.

In addition to their main purpose, the filters provide the ancillary benefit of a latent-variable analysis of the noise they filter out: the noise matrix is constructed from principal components defined by scores and loadings, in analogy to the correlated part of the model. The introduction of OSC also changes the dependence of the model. Since $\boldsymbol{X}$ and $\boldsymbol{Y}$ are both filtered using the same technique, they are now of equal importance, and two equally valid regressions are established between them. The symmetry between the data blocks allows O2-PLS to apply without hesitation to situations—*e.g.* correlations between serum and tissue cytokine levels—in which either data block could conceivably be considered the independent variable set. O2-PLS's features combine to produce a model with more elements, but even greater flexibility and interpretability with less cost and complexity than variable selection, and additional insight:

x decomposition:	$\boldsymbol{X} = \boldsymbol{T}\boldsymbol{P}^T + \boldsymbol{X}_{Y_o} + \boldsymbol{E}$	(1.15)
y decomposition:	$\boldsymbol{Y} = \boldsymbol{U}\boldsymbol{Q}^T + \boldsymbol{Y}_{X_o} + \boldsymbol{F}$	(1.16)
Orthogonal decomposition of $\boldsymbol{X}$:	$\boldsymbol{X}_{Y_o} = \boldsymbol{T}_{Y_o}{\boldsymbol{P}_{Y_o}}^T$	(1.17)
Orthogonal decomposition of $\boldsymbol{Y}$:	$\boldsymbol{Y}_{X_o} = \boldsymbol{U}_{X_o}{\boldsymbol{Q}_{X_o}}^T$	(1.18)
Predictive regression of $\boldsymbol{X}$ onto $\boldsymbol{Y}$:	$\boldsymbol{T} = \boldsymbol{U}\boldsymbol{B}_{U\to T} + \boldsymbol{H}_{U\to T}$	(1.19)
Predictive regression of $\boldsymbol{Y}$ onto $\boldsymbol{X}$:	$\boldsymbol{U} = \boldsymbol{T}\boldsymbol{B}_{T\to U} + \boldsymbol{H}_{T\to U}$	(1.20)

Note that, in this situation, the predictive regression is symmetric and established between the scores, rather than between $\boldsymbol{X}$ and $\boldsymbol{Y}$ directly. Thus, if a model is already established, in order to predict the model's estimates for new data, the orthogonal components must be filtered out before the estimates are calculated.

In analogy with SIMPLS, the O2-PLS algorithm is based on a singular value decomposition of the correlation matrix $\boldsymbol{K}$, but rather than consistent deflation of $\boldsymbol{K}$, $\boldsymbol{X}$ and $\boldsymbol{Y}$ are deflated themselves using their projections onto a vector space orthogonal to the principal components of $\boldsymbol{K}$. At the end, the deflations are augmented to form the orthogonal projections. An outline of the full procedure is described in Algorithms 2 and3. As input, it takes the data matrices $\boldsymbol{X}$ and $\boldsymbol{Y}$, along with the number of predictive, X-orthogonal, and Y-orthogonal components to keep.

O2-PLS can be combined with other variable selection techniques, but because of the OSC filter, this is rarely necessary. Instead, the final model is chosen by picking optimal component numbers a, a_{X_o}, and a_{Y_o} by minimizing a cross-validation error, by a visual scree plot technique [57, 58], or by some combination of the two. The outcome is a powerful, informative, flexible, and expressive model that captures and analyzes a wide range of variation in the data to provide useful insights.

Algorithm 2 O2-PLS, Part 1

Require: Data matrices $\boldsymbol{X}$, $\boldsymbol{Y}$, and maximum dimension numbers a, a_{Y_o}, and a_{X_o}.
Ensure: The matrices $\boldsymbol{P}, \boldsymbol{Q}, \boldsymbol{T}, \boldsymbol{U}, \boldsymbol{P}_{Y_o}, \boldsymbol{Q}_{X_o}, \boldsymbol{T}_{Y_o}, \boldsymbol{U}_{X_o}, \boldsymbol{B}_{U\to T}$, and $\boldsymbol{B}_{T\to U}$ fully specify an O2-PLS model of Y vs X.

$\boldsymbol{K} \leftarrow \boldsymbol{X}^T\boldsymbol{Y}$
$\boldsymbol{W}\boldsymbol{S}\boldsymbol{C}^T = \boldsymbol{K}$ ▷ Compact SVD of $\boldsymbol{K}$

$\boldsymbol{W} \leftarrow \boldsymbol{W}\left(\boldsymbol{W}^T\boldsymbol{W}\right)^{-\frac{1}{2}}$ ▷ Normalize $\boldsymbol{W}, \boldsymbol{C}$
$\boldsymbol{C} \leftarrow \boldsymbol{C}\left(\boldsymbol{C}^T\boldsymbol{C}\right)^{-\frac{1}{2}}$
$\boldsymbol{W} \leftarrow$ the first a columns of $\boldsymbol{W}$
$\boldsymbol{C} \leftarrow$ the first a columns of $\boldsymbol{C}$

for all $i \in \{1, \ldots, a_{Y_o}\}$ **do** ▷ OSC on $\boldsymbol{X}$
  $\boldsymbol{T} \leftarrow \boldsymbol{X}\boldsymbol{W}$
  $\boldsymbol{P} \leftarrow \boldsymbol{X}^T\boldsymbol{T}\left(\boldsymbol{T}^T\boldsymbol{T}\right)^{-1}$

  $\boldsymbol{K}_{Y_o} \leftarrow \boldsymbol{P} - \boldsymbol{W}\boldsymbol{W}^T\boldsymbol{P}$ ▷ $\boldsymbol{P}$, orthogonalized to $\boldsymbol{W}$: heart of OSC

  Subtract out most significant component of $\boldsymbol{K}_{Y_o}$ from $\boldsymbol{X}$:
  $\boldsymbol{W}_{Y_o}\boldsymbol{S}_{Y_o}\boldsymbol{C}_{Y_o}^T = \boldsymbol{K}_{Y_o}$ ▷ Compact SVD of $\boldsymbol{K}_{Y_o}$

  $\boldsymbol{w}_{Y_o}^{(i)} \leftarrow$ the first column of $\boldsymbol{W}_{Y_o}$
  $\boldsymbol{t}_{Y_o}^{(i)} \leftarrow \boldsymbol{X}\boldsymbol{w}_{Y_o}^{(i)}$ ▷ x-scores for orthogonal component i
  $\boldsymbol{t}_{Y_o}^{(i)} \leftarrow \boldsymbol{t}_{Y_o}^{(i)} / \left(\boldsymbol{t}_{Y_o}^{(i)^T}\boldsymbol{t}_{Y_o}^{(i)}\right)$ ▷ Normalize $\boldsymbol{t}_{Y_o}^{(i)}$
  $\boldsymbol{p}_{Y_o}^{(i)} \leftarrow \boldsymbol{X}^T\boldsymbol{t}_{Y_o}^{(i)}$ ▷ x-loadings for component i: a least-squares fit of $\boldsymbol{X}$ to $\boldsymbol{t}_{Y_o}^{(i)}$

  $\boldsymbol{X} \leftarrow \boldsymbol{X} - \boldsymbol{t}_{Y_o}^{(i)}\boldsymbol{p}_{Y_o}^{(i)^T}$ ▷ Deflate the orthogonal component out of $\boldsymbol{X}$
end for

Algorithm 3 O2-PLS, Part 2

for all $i \in \{1, \ldots, a_{X_o}\}$ **do** ▷ OSC on $\boldsymbol{Y}$

$\boldsymbol{U} \leftarrow \boldsymbol{Y}\boldsymbol{C}$

$\boldsymbol{Q} \leftarrow \boldsymbol{Y}^T\boldsymbol{U}\left(\boldsymbol{U}^T\boldsymbol{U}\right)^{-1}$

$\boldsymbol{K}_{X_o} \leftarrow \boldsymbol{Q} - \boldsymbol{C}\boldsymbol{C}^T\boldsymbol{Q}$ ▷ $\boldsymbol{Q}$, orthogonalized to $\boldsymbol{C}$: heart of OSC

Subtract out most significant component of $\boldsymbol{K}_{X_o}$ from $\boldsymbol{Y}$:

$\boldsymbol{C}_{X_o}\boldsymbol{S}_{X_o}\boldsymbol{W}_{X_o}^T = \boldsymbol{K}_{X_o}$ ▷ Compact SVD of $\boldsymbol{K}_{X_o}$

$\boldsymbol{c}_{X_o}^{(i)} \leftarrow$ the first column of $\boldsymbol{C}_{X_o}$

$\boldsymbol{u}_{X_o}^{(i)} \leftarrow \boldsymbol{Y}\boldsymbol{c}_{X_o}^{(i)}$ ▷ y-scores for orthogonal component i

$\boldsymbol{u}_{X_o}^{(i)} \leftarrow \boldsymbol{u}_{X_o}^{(i)} / \left(\boldsymbol{u}_{X_o}^{(i)\,T}\boldsymbol{u}_{X_o}^{(i)}\right)$ ▷ Normalize $\boldsymbol{u}_{X_o}^{(i)}$

$\boldsymbol{q}_{X_o}^{(i)} \leftarrow \boldsymbol{Y}^T\boldsymbol{u}_{X_o}^{(i)}$ ▷ y-loadings for component i: a least-squares fit of $\boldsymbol{Y}$ to $\boldsymbol{u}_{X_o}^{(i)}$

$\boldsymbol{Y} \leftarrow \boldsymbol{Y} - \boldsymbol{u}_{X_o}^{(i)}\boldsymbol{q}_{X_o}^{(i)\,T}$ ▷ Deflate the orthogonal component out of $\boldsymbol{Y}$

end for

$\boldsymbol{T} \leftarrow \boldsymbol{X}\boldsymbol{W}$ ▷ Final reëvaluation of the scores and loadings

$\boldsymbol{P} \leftarrow \boldsymbol{X}^T\boldsymbol{T}\left(\boldsymbol{T}^T\boldsymbol{T}\right)^{-1}$

$\boldsymbol{U} \leftarrow \boldsymbol{Y}\boldsymbol{C}$

$\boldsymbol{Q} \leftarrow \boldsymbol{Y}^T\boldsymbol{U}\left(\boldsymbol{U}^T\boldsymbol{U}\right)^{-1}$

▷ Combine orthogonal components into matrices

$\boldsymbol{W}_{Y_o} \leftarrow \left[\boldsymbol{w}_{Y_o}^{(1)}|\ldots|\boldsymbol{w}_{Y_o}^{(a_{Y_o})}\right]$

$\boldsymbol{P}_{Y_o} \leftarrow \left[\boldsymbol{p}_{Y_o}^{(1)}|\ldots|\boldsymbol{p}_{Y_o}^{(a_{Y_o})}\right]$

$\boldsymbol{Q}_{Y_o} \leftarrow \left[\boldsymbol{q}_{Y_o}^{(1)}|\ldots|\boldsymbol{q}_{Y_o}^{(a_{Y_o})}\right]$

$\boldsymbol{C}_{X_o} \leftarrow \left[\boldsymbol{c}_{X_o}^{(1)}|\ldots|\boldsymbol{c}_{X_o}^{(a_{X_o})}\right]$

$\boldsymbol{T}_{X_o} \leftarrow \left[\boldsymbol{t}_{X_o}^{(1)}|\ldots|\boldsymbol{t}_{X_o}^{(a_{X_o})}\right]$

$\boldsymbol{U}_{X_o} \leftarrow \left[\boldsymbol{u}_{X_o}^{(1)}|\ldots|\boldsymbol{u}_{X_o}^{(a_{X_o})}\right]$

$\boldsymbol{B}_{U\to T} = \left(\boldsymbol{U}^T\boldsymbol{U}\right)^{-1}\boldsymbol{U}^T\boldsymbol{T}$ ▷ Regressions are least-squares fits of the scores

$\boldsymbol{B}_{T\to U} = \left(\boldsymbol{T}^T\boldsymbol{T}\right)^{-1}\boldsymbol{T}^T\boldsymbol{U}$

1.4.6 O*n*-PLS and future directions

Finally, the ideas developed in O2-PLS and O-PLS can be extended even further. The flexible nature of O2-PLS, combined with its completely symmetric treatment of both the X and Y data blocks leads to a direct extension of the method to any number of blocks: O*n*-PLS [59]. In O*n*-PLS, each block is decomposed into a predictive part, which takes into account correlations with every other block, and an orthogonal part, defined relative to the predictive part, just as in O2-PLS. O*n*-PLS has not yet been widely used, due to its novelty, but it offers a promising new direction for analysis of multi-block biological data. Finally, post-processing approaches similar to ones developed for O-PLS [60–62] may be useful extensions of O*n*-PLS as well.

Chapter 2

Molecular dynamics studies of the alloreactive T cell response

2.1 Summary

T cells orchestrate adaptive, pathogen-specific immune responses. T cells have a surface receptor (called TCR) whose ligands are complexes (pMHCs) of peptides (derived from pathogens or host proteins) and major histocompatibility complex proteins (MHCs). MHC proteins vary between hosts. During organ transplants, host TCRs interact with peptides present in complex with genetically different MHCs. This usually causes a vigorous immune response—alloreactivity. Studies of alloreactive protein interactions have yielded results that present a puzzle. Some crystallographic studies concluded that the alloreactive TCR/MHC interface is essentially unaffected by changing the TCR peptide-binding region, suggesting that

the peptide does not influence the interface. Another biochemical study concluded from mutation data that different peptides can alter the binding interface with the same TCR. To explore the origin of this puzzle, we used molecular dynamics simulations to study the dependence of the TCR/pMHC interface on changes in both the peptide and the TCR. Our simulations show that the footprint of the TCR on the pMHC is insensitive to mutations of the TCR peptide-binding loops, but peptide mutations can make multiple local changes to TCR/pMHC contacts. Therefore, our results demonstrate that the structural and mutation data do not conflict and reveal how subtle, but important, characteristics of the alloreactive TCR/pMHC interface are influenced by the TCR and the peptide.

2.2 Introduction

The adaptive immune system enables higher organisms, like humans, to protect themselves with pathogen-specific responses against a diverse and evolving world of microbes. T lymphocytes (T cells) are key orchestrators of the adaptive immune response. To perform their functions, they must be activated. Activation is predicated on sufficiently strong binding of a T cell's antigen receptor (T cell receptor, TCR) to a ligand. The ligand consists of a short peptide fragment held in the cleft of a membrane-bound major histocompatibility complex protein (MHC) displayed on the surface of an antigen-presenting cell. T cell activation can lead to a variety of effector immune functions.

Immature T cells undergo development in the thymus, where they interact with pMHC complexes derived from the host proteome. To survive elimination during the development process, the T cells must not interact too strongly with any of these self-pMHC complexes (negative selection), but must bind with sufficient affinity to at least one pMHC (positive selection) [63–70]. This selection process largely

inhibits autoimmune T cells from joining the immune system and ensures that the surviving T cells can recognize foreign peptides presented on the host's own MHC molecules with extraordinary specificity [71,72]. Because of thymic selection, peptides derived from the hosts' own proteins do not produce a strong interaction, but foreign-derived peptides do.

Sometimes, such as during organ transplantation, mature T cells encounter pMHC complexes on cells from a genetically different (allogeneic) member of the same species. Since MHC genes are highly polymorphic, allogeneic pMHCs (allo-pMHCs) present previously unseen MHC surfaces to the TCRs. Furthermore, since the most variable regions of the MHC occur along the peptide binding cleft, the peptides presented by allo-MHCs likely differ in sequence and conformation from the self-peptides used to train the TCR in the thymus, even though they originate from the same proteins. Up to 10% of the T cell repertoire can cross-react with any particular pMHC on the allogeneic cells—1000 times as many T cells as the 0.01% of the repertoire activated during the response to a virus [73–81]. This intense response, known as alloreactivity, makes organ transplantation impossible without immuno-suppression.

Much experimental work has been dedicated to elucidating the roles and relative importance of the peptide and MHC in alloreactivity, and a large number of these studies [84–91] have examined the interaction footprint—the interface between the TCR and the set of pMHC residues that come into contact with it, which includes the peptide and the α_1 and α_2 helices of the MHC (Figure 2-1). A question of particular interest, has been the energetic and structural impact of the peptide on an allo-pMHC/TCR footprint. The question has been actively explored by biochemical mutation experiments and X-ray crystallography.

Some biochemical experiments have studied how a TCR interacts with different peptides in the same allo-MHC. But these studies do not provide direct structural

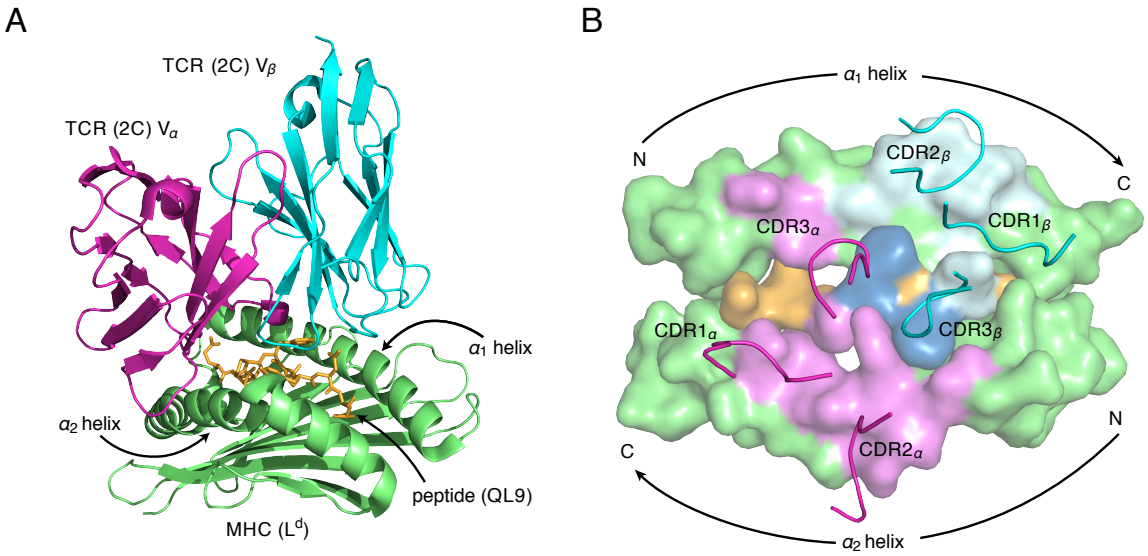


Figure 2-1: The 2C TCR binds strongly to the allo-MHC L^d, contacting a "footprint" set of residues on the pMHC. (a) A crystal structure (PDB 2019) of the variable chains of the 2C TCR in complex with the peptide QL9 and the MHC L^d shows the relative orientation of TCR on MHC [82]. (b) The diagram of the TCR/pMHC footprint highlights L^d residues that are in contact with 2C. A pMHC residue is considered to be in contact with TCR if it has a non-hydrogen atom within 4.5 Å of a non-hygrogen atom in a TCR residue. pMHC residues that do contact TCR are shaded with the color of the chain(s) that they contact: magenta for V_α, cyan for V_β, and dark blue for both. Above the MHC, the 2C CDR loops are indicated in the color of the chain they belong to: magenta for V_α, cyan for V_β. The directionality of the protein backbone is indicated by arrows that point from the amino terminus to the carboxy terminus. Molecular visualizations for this figure were created with PyMol [83].

data [92, 93]. One study, by Felix et al., concluded that peptide mutations could have a noticeable impact on the TCR/pMHC footprint by affecting contacts between MHC and TCR [92]. In the study, the authors mutated several residues along the α-helices of a particular allo-MHC and observed the effects of those mutations on T cell activation, in order to assess which residues were likely to contact the TCR. The authors found that when different peptides were bound by the same allo-MHC, different subsets of MHC residues impacted recognition, implying that the peptides affected the TCR/MHC contacts.

Since the first complete TCR/pMHC crystal structure in 1996 [18], crystal structures have also been used to study several TCR/pMHC footprints. Many of these footprints have been between TCR and peptide/self-MHC complexes [18,94–99], but structural studies have dealt with alloreactive complexes as well [19,82,100–107]. Two of these studies [19, 107] have made direct structural comparisons of the effects of different peptides of varying affinity on the TCR/allo-pMHC interface. A recent crystallographic study by Colf et al. concluded that the peptide has little impact on the footprint. The authors reached this conclusion by showing that mutations on the CDR3$_\alpha$ peptide-recognition loop of a TCR bound to an allo-pMHC complex did not impact the structure, despite increasing the binding affinity by over two orders of magnitude [82]. Importantly, however, the study did not involve mutation of the peptide itself. A more physical and chemical understanding of the intermolecular interactions involved at the TCR/pMHC interface could shed light on these seemingly conflicting biochemical and crystallographic results. Such insight could also motivate further experiments.

One direct way to compare the two results is to understand a single system that encompasses *both* peptide mutation and structural information. To this end, we have performed an *in silico* analog of the peptide mutation experiments designed by Felix et al. on TCR/pMHC structures obtained by Colf et al. [82]. Using molecular

dynamics simulations [17], we analyzed atomistic models of TCR/allo-pMHC complexes while independently changing both the TCR and the peptide, which allowed us to directly compare the effects of each kind of mutation on the TCR/pMHC interface. In our simulations, as in the crystallographic study, mutation of the CDR3$_\alpha$ loop of the TCR did not induce a significant change in the TCR/allo-pMHC footprint, compared to the significant differences between the TCR/allo-pMHC and TCR/self-pMHC footprints. However, our simulations also showed that certain *peptide* mutations *can* affect the TCR/pMHC interface. These peptide mutations not only affected the peptide-TCR contacts, but also influenced which MHC residues came into contact with TCR, even though they did not induce a change in the overall orientation of TCR on MHC—a finding that confirms the conclusions of Felix et al. [92]. Our simulations thus demonstrate that the crystallographic results and biochemical results are not in conflict, because mutations to the CDR3$_\alpha$ loop are not necessarily equivalent to peptide mutations. Our results highlight the potential of the peptide to impact the TCR/pMHC interface by making local contact changes and suggest that detailed chemical interactions at the interface between the TCR, peptide, and MHC can all play a part in the ultimate structure of the alloreactive TCR/pMHC interface.

Our findings are consistent with an attractive model for TCR/pMHC interactions in which the TCR docks over the pMHC and scans the ligand for a sufficient number of interactions that confer the TCR/pMHC complex a sufficient lifetime [12, 70, 108, 109]. If this necessary condition for recognition is met, structural rearrangements occur to acquire enhanced affinity. The specific character of these relatively modest rearrangements [91] depends on the particular TCR/pMHC pair under consideration.

2.3 Methods

2.3.1 Structure preparation

The X-ray structures of the C9X and M9X [82] complexes (PDB IDs 2O19 and 2E7L, respectively) were used for the initial coordinates of all calculations. Three mutant-peptide variants were generated from each crystal structure by removing the atoms that corresponded to the mutation (Table 2.1). This resulted in a total of eight systems to simulate. Hydrogen atoms were introduced into the structures with a stereochemical algorithm [110]. The protonation states of titratable residues were assigned based on visual inspection and prior results for similar systems [28–30]. The structures were minimized with constraints on heavy atom positions and solvated with explicit TIP3P [23, 111] water molecules in 89 Å rhombic dodecahedra with periodic boundary conditions. 39 K^+ ions and 37 – 38 Cl^- ions were added to each system to neutralize the total charge and simulate a 0.15 M KCl concentration. The ions were placed with random initial coordinates, then subjected to 2000 steps of a Metropolis Monte Carlo simulation in order to equilibrate their positions. After solvation, hydrogens, solvent atoms, and protein atoms were all minimized with restraints in stages. First, hydrogens were minimized for 1000 steps without restraints, while heavy atoms were fixed. Then, only the solvent atoms were subjected to 1000 additional steps of constrained minimization, with force constants of 0.5 kcal/mol/Å^2, while protein atoms were fixed. After that, all atoms were unfixed, a 0.1 kcal/mol/Å^2 harmonic restraint was placed on the protein atoms, and the system was subjected to 1000 more steps of minimization. Finally, the entire system was minimized again for 1000 steps without restraints. The average number of atoms per system was 47,355.

2.3.2 Molecular dynamics simulations

All simulations were performed using the CHARMM [15, 112] molecular dynamics program (version c34b1) with the CHARMM 22 [23] force field and the CMAP correction for the peptide backbone dihedrals [113]. The simulation protocol was influenced by previous TCR/pMHC simulations [28–30]. Molecular dynamics simulations were performed with holomorphic constraints on the hydrogen atoms [114], which allowed the use of 2 femtosecond integration time steps. Non-bonded van der Waals interactions were truncated at 9 Å using a force-switching algorithm [115]. Electrostatic terms were calculated using the Particle-Mesh Ewald summation method [116], and the real-space terms were evaluated with a 9 Å cutoff. After solvation and the energy minimizations described above, each system was heated and equilibrated in several stages. First, the entire system was heated from 10 K to 300 K over 100 ps without restraints. Then a 3.0 kcal/mol/Å^2 harmonic restraint was placed on the heavy protein atoms, and the ions and solvent were equilibrated with three repetitions of a heating and cooling cycle. The cycle ran constant-temperature dynamics for 50 ps at 300 K, then heated the system to 450 K, ran constant-temperature dynamics for 100 ps, cooled the system down to 300 K, then ran another 100 ps of constant-temperature dynamics at 300 K. After the solvent equilibration, a mass-weighted 0.75 kcal/mol/Å^2 harmonic restraint was placed on protein heavy atoms and the system was equilibrated for 100 ps with constant-temperature dynamics. The restrained dynamics were followed with 100 ps of unrestrained constant-temperature dynamics, and 200 ps of unrestrained constant-temperature, constant-pressure dynamics. Simulations were then continued for 15 ns. Throughout each simulation, constant temperature was maintained with the Nosé-Hoover thermostat [117, 118], and constant pressure was maintained with the Langevin piston method [119].

Table 2.1: Lower-affinity peptide mutants are good candidates for simulation.

Peptide[a]	Sequence	TCR/pMHC $K_a\left(M^{-1}\right)$[b]	Reference
QL9	QLSPFPFDL	$1.0 - 2.0 \cdot 10^7$	[120]
QL9-A6	QLSPFAFDL	$2.0 \cdot 10^5$	[121]
P2CA	-LSPFPFDL	$2.0 \cdot 10^6$	[120]
P2CA-A5	-LSPFAFDL	$1.6 \cdot 10^4$	[120]

[a] In order to assess the impact of peptide mutations on the TCR/pMHC footprint, we chose to simulate the 2C/QL9-L^d system with several lower-affinity peptide variants. The table lists sequences and binding affinities of these peptides. The mutants were chosen to impact the binding affinity by at least a factor of 5, so as to not simulate "null" mutations.

[b] Binding affinities were obtained from solution measurements of peptide on L^d, presented to 2C.

2.4 Results

2.4.1 Alloreactive model

The alloreactive model used in our simulations is based on the murine 2C TCR. T cells containing this TCR were found in a mouse—whose own MHC molecules expressed the K^b-haplotype—injected with cells whose MHCs had the L^dhaplotype [1, 122]. Thus, 2C TCRs recognize foreign and self-peptides on the K^b self-MHC and also bind strongly to peptides presented on the L^d allo-MHC (Figure 2-1). Crystallographic experiments by Colf et al. [82] have obtained the structure of the variable part of a 2C TCR in contact with a peptide (QL9; see Table 2.1 for sequence) held in the cleft of a stabilized portion of L^d (α_1/α_2 domains, residues 1 – 180 of the heavy chain). The crystal structure of the same ligand bound to m6, a mutant of 2C with even higher affinity for QL9/L^d, was also obtained. The only difference in sequence between 2C and m6 occurs in residues 94 – 103 of the CDR3$_\alpha$ loop, which is changed from the sequence GFASA in 2C to the sequence HQGRY in m6. Despite the 100-fold change in affinity for the ligand caused by this mutation, it

had no significant impact on the TCR/pMHC footprint, especially when compared to the structural difference between the 2C/QL9-L^d structure and previously-obtained 2C/K^b self-pMHC structures [82, 94]. This result led Colf et al. to conclude that the peptide did not play an important structural role in the footprint because, even though interactions between the peptide and TCR had significantly changed, the footprint had not.

2.4.2 Molecular dynamics simulations

To study the effect of both TCR and peptide mutation on the TCR/pMHC footprint, we performed molecular dynamics simulations on a total of eight systems. For the peptide mutants, we used four peptides (Table 2.1), including QL9. The three additional peptides were derived from QL9 by independently introducing two mutations: deletion of the terminal glutamine, and mutation of the carboxy-proximal proline to an alanine. These mutations were chosen for their ability to noticeably affect the binding affinity of the peptide for 2C [120, 121]. Each of the four pMHCs was simulated with both 2C and m6 TCR, with the m6 simulations intended to test the effect of TCR mutation. The simulated systems are summarized in Table 2.2, along with their identifying abbreviations.

Since the peptide mutations involved only removal of atoms, the initial coordinates of the mutant peptides were the same as the crystal coordinates of QL9, except for the deleted atoms. The molecular dynamics simulations were carried out for 15 ns, and coordinates were saved at every 2 ps for analysis. Energetic and structural properties of the systems were calculated by averaging over the last 5 ns of the trajectories, after the systems had reached equilibrated states. Equilibration was determined by a consistent plateau in the RMSD (root-mean-square deviations) of the backbone coordinates from the crystal structure (Supplemental Figure A-1). The binding free energy of the m6 TCR for pMHC was calculated

Table 2.2: For clarity and conciseness, we use abbreviations to refer to the TCR/pMHC systems discussed in this article.

Peptide	TCR	MHC	Source	Abbrev.[a]
QL9	2C	L^d	dynamics	C9P
	m6	L^d	dynamics	M9P
	2C	L^d	structure	C9X
	m6	L^d	structure	M9X
QL9-A6	2C	L^d	dynamics	C9A
	m6	L^d	dynamics	M9A
p2Ca	2C	L^d	dynamics	C8P
	m6	L^d	dynamics	M8P
p2Ca-A5	2C	L^d	dynamics	C8A
	m6	L^d	dynamics	M8A
SIYR	2C	K^b	dynamics	CSK
	2C	K^b	structure	CSX

[a] The abbreviation code describes the distinguishing features of a system with three characters. For the L^d simulations, the first character describes the TCR involved: "C" for 2C, "M" for m6. The second character describes how many amino acids are in the peptide: "9" for QL9 and its mutants, "8" for p2Ca and its mutants. The third character describes the mutant variation if the abbreviation refers to the simulated structure: "P" for the original amino acids with proline at position 5 or 6, QL9 and p2Ca, "A" for the alanine-mutants QL9-A6 and p2Ca-A5. The third character can also be "X," which is a special case indicating that the abbreviation refers to a crystal structure. Finally, the K^b systems are named in a manner to mimic consistency with the L^d systems: "C" for 2C, "S" for the SIYR peptide, "K" for K^b, and "X" for the crystal structure.

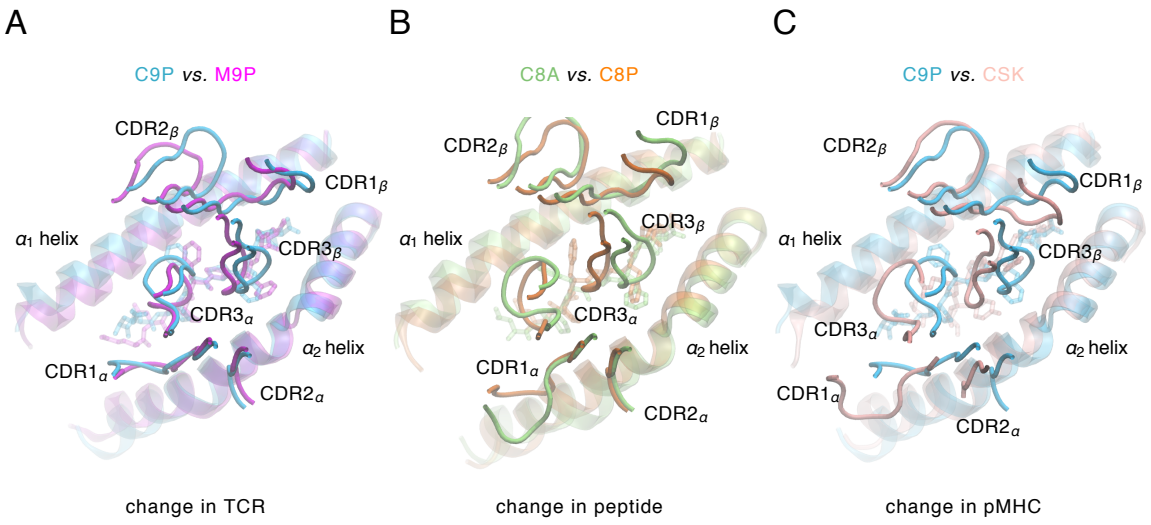


Figure 2-2: Average structures of the TCR/pMHC footprints compare the effects of TCR, peptide, and pMHC changes. The comparison highlights the ability of the simulations to reproduce qualitative experimental results. The average structures are aligned via their MHC backbone atoms to allow comparisons between TCR loop displacement and TCR orientation. The differences between the two structures are most noticeable in (c), and least pronounced in (a), in agreement with experimental results. (b), which highlights differences from peptide mutation, shows differences that lie in between the two extremes. The average structures were obtained from the 5 ns of the trajectories, with rotation and translation removed. Molecular visualizations for this figure were created with VMD [123].

to be stronger (more negative) than that of 2C (Supplemental Table A.1), which corresponds qualitatively to experimental binding affinity measurements [104].

To provide more points of reference, a model of the 2C/K^b system was also simulated. The resolution of the crystal structure referenced by Colf et al., originally obtained by Garcia et al. [94], was not high enough for consistent simulations. Instead, a related structure, which replaced the original DEV8 self-peptide with the superagonist SIYR peptide, was used [96]. This structure had a higher resolution, yet differed little (RMSD 0.92 Å for C_α atoms) from the original 2C/K^b/DEV8 structure and could therefore serve as an approximate model.

2.4.3 Average structures from the dynamics highlight the overall effect of peptide mutation

In order to evaluate the qualitative effects of peptide mutation, we calculated average TCR/pMHC structures from each trajectory by taking the mean position of each atom during the last 5 ns of the trajectory. Figure 2-2 displays some of these average structures superimposed on one another for comparison. The structures in each part of the figure are aligned along their MHC backbone atoms to highlight differences in TCR loop position and TCR/MHC orientation. Figure 2-2 is designed to show the effect of TCR mutation (Figure 2-2a), the effect of peptide mutation (Figure 2-2b), and the effect of switching from self-pMHC to allo-pMHC (Figure 2-2c). Looking from Figure 2-2a to Figure 2-2c, the differences between the structures in each pair increase. In Figure 2-2a, the qualitative features Colf et al. found are reproduced: the overall orientation of both TCRs is nearly identical, [82] despite a few minor loop rearrangements, leading to an average backbone RMSD of 2.7 Å for the loops. By stark contrast, in Figure 2-2c, almost all of the TCR loops have significantly different average positions, with an average backbone RMSD of 3.5 Å [91], and it is clear that 2C binds to the different pMHCs at different angles. Figure 2-2a and Figure 2-2c together demonstrate that our simulations retain qualitative similarity to the crystallographers' finding that CDR3 loop mutation is insignificant, especially compared to the differences between self- and allo-pMHC [82]. Figure 2-2b shows how the effects of peptide mutation compare with the two "extremes" of Figure 2-2a (no effect) and Figure 2-2c (drastic effect). The two peptide mutant systems shown in this subfigure, C8A and C8P, exhibit the greatest difference from one another among all choices of peptide mutants for a given TCR, highlighting the maximum observed extent of peptide mutation. Comparing C8A to C8P shows that, although there is no global rearrangement of binding orientation, as in Figure 2-2c, there are many differences between the conforma-

tions of the 2C CDR loops, and the differences are distributed throughout most of the CDR loops, yielding an average backbone RMSD of 3.4 Å. Every loop except CDR2$_\alpha$ exhibits a noticeable difference between the two structures. These unique loop conformations lead to significantly different TCR/pMHC contacts. In particular, the most significant differences are seen in the CDR3 loops, which contact the peptide directly. The peptides, however, are not adopting vastly different conformations (backbone RMSD 1.9 Å). This result suggests that the peptides interact with the TCR in a highly dissimilar manner despite their relatively similar orientation and that mutation can affect the TCR/pMHC contacts through a complex interplay of altered interactions between the TCR and the bound pMHC.

2.4.4 TCR/pMHC contact distributions allow quantitative comparison of the effects of mutation

To measure the impact of peptide mutations on the footprint, we computed the number of contacts that each pMHC residue makes with TCR and averaged these quantities over the structures obtained during the last 5 ns of simulation. We defined any two residues to be in contact when a non-hydrogen atom in one residue was within 4.5 Å of a non-hydrogen atom in the other residue. This average-contact approach qualitatively differs from analyzing any single "representative" structure taken from the simulation trajectory, since a single structure may not accurately represent the overall ensemble of structures well enough, especially when examining the contact footprint. Thus, an analysis based on a single structure would likely suffer from ambiguities in interpreting the results.

The result of the average contact calculations, shown in Figure 2-3 and Figure 2-4, is a picture of the interaction footprints as functions of pMHC sequence. In each plot, the ordinate lists the positional indices and amino acid abbreviations for the residues that constitute a particular segment of the pMHC—either the α_1 helix

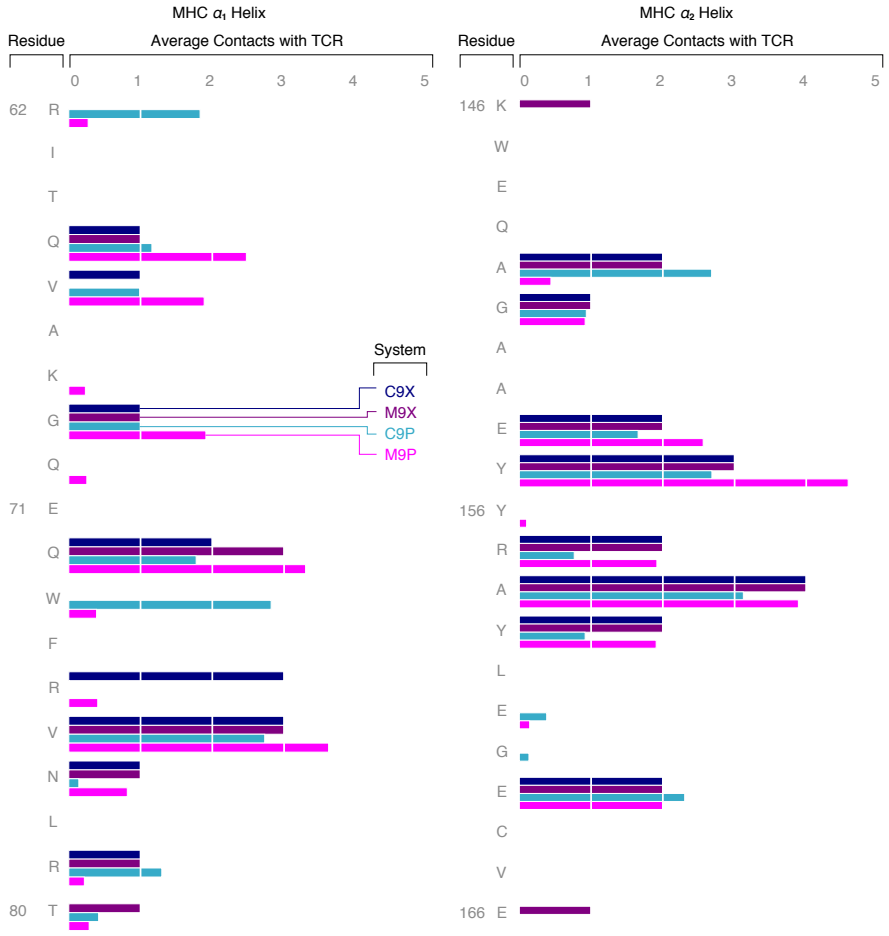


Figure 2-3: MHC/TCR contact distributions show a quantitative description of the TCR/pMHC footprint. The figure compares the crystal structure footprints to their corresponding simulations in order to establish the accuracy of the simulations. In general, the simulations reproduce the result that mutation of the TCR does not make widespread changes to the footprint. The ordinate axes list the residue index and single-letter amino acid abbreviation for the residues of the MHC α_1 and α_2 helices. For the simulations (C9P and M9P), the length of each bar represents how many contacts, on average, that MHC residue made with any TCR residue during the last 5 ns of the dynamics trajectory. For the crystal structure data (C9X and M9X), the length of a bar represents the exact number of contacts observed in that structure. Two residues were defined to be in contact if a non-hydrogen atom in one residue was within 4.5 Å of a non-hydrogen atom within another. S.e.m. error for each bar length was calculated but is too small to be visible.

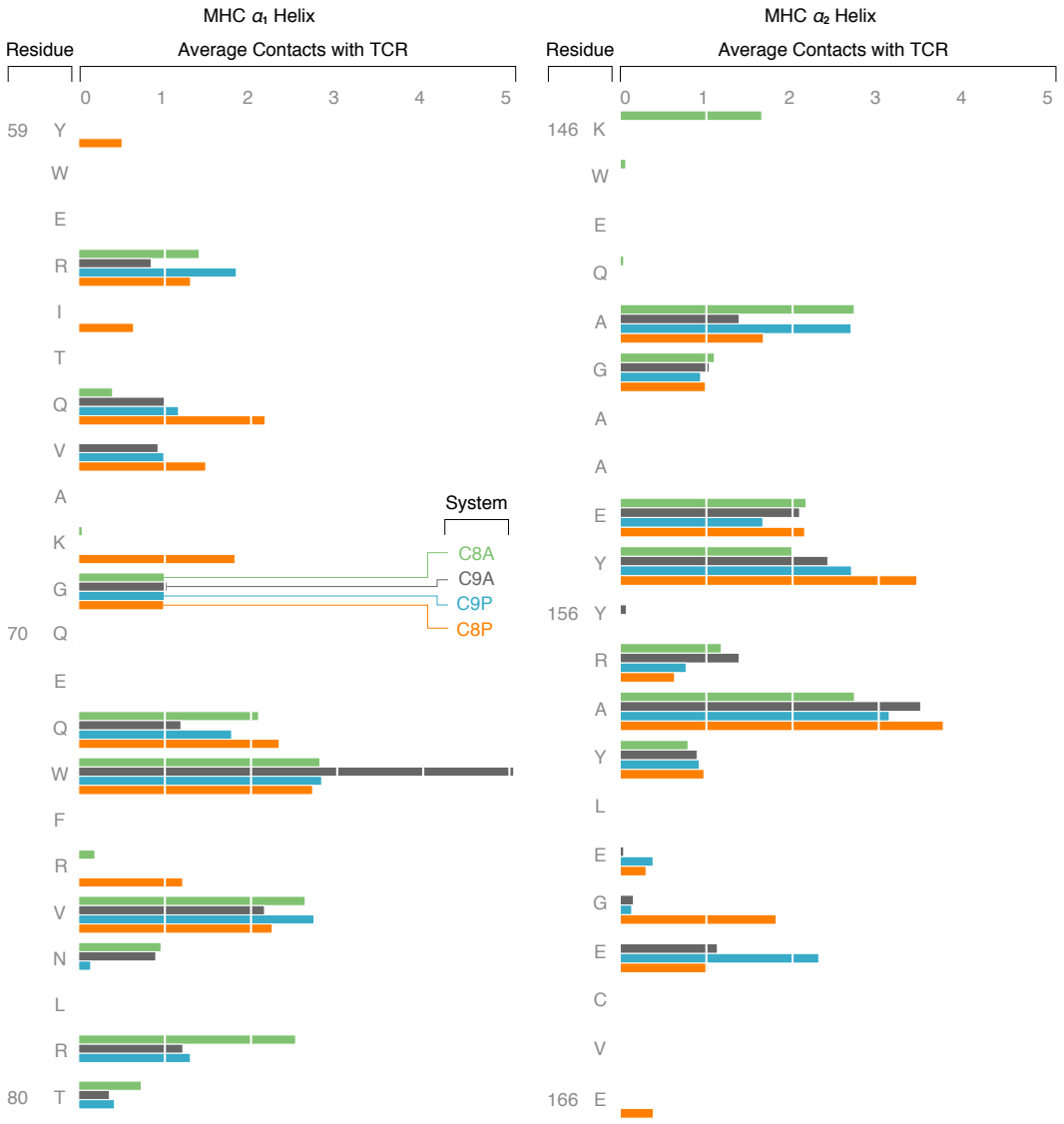


Figure 2-4: Peptide mutation can induce local TCR/MHC contact changes. MHC/TCR contact distributions of peptide-mutant systems show the effects of peptide mutation on the interface. Some of these effects are significant: the C8A distribution is noticeably different from the C8P distribution. The ordinate axes list the residue index and single-letter amino acid abbreviation for the residues of the MHC α_1 and α_2 helices. The length of each bar represents how many contacts, on average, that MHC residue made with any TCR residue during the last 5 ns of the dynamics trajectory. Contacts were defined the same way as Figure 2-3. S.e.m. error for each bar length was calculated but is too small to be visible.

or the α_2 helix. Around each point, which represents an individual pMHC residue, a group of bars is clustered. The set of all bars of a single color represents a particular TCR/pMHC system, as described by the legend and Table 2.2, and the length of a bar represents the average number of contacts that a pMHC residue made with TCR. Figure 2-4 shows the effect of peptide mutation on the 2C footprint by comparing all simulated systems that involve the binding of 2C TCR to different peptides: C9P, C9A, C8P, and C8A, respectively. Figure 2-3 serves as a control by comparing the footprints of the 2C and M6 crystal structures (C9X, M9X) with their corresponding simulations (C9P, M9P). The control set provides two methods of testing the simulations against experimental results, showing (1) how well the simulations reproduce the experimental conclusions that mutation from 2C to M6 does not have a significant effect on the footprint, and (2) how similar the footprints generated by the simulations are to their experimental counterparts. To be numerically significant, any changes observed from peptide mutation would have to be larger than the differences between crystal structures and their simulation counterparts, and also larger than the differences between the 2C and M6 TCR simulations.

Figure 2-4 directly shows how peptide mutation changes local contacts between MHC residues and TCR in several locations throughout the footprint, in direct support of the findings of Felix et al. [92], that peptide mutations can rearrange the finely-paired contacts between TCR and MHC. For a more quantitative comparison of the footprints, we present a single system's footprint (the set of all bars of the same color) as a discrete "distribution" of contacts over the sequence of the pMHC. The distributions of different systems can be compared segment-by-segment by their "means," $\bar{r}_c$, shown in Figure 2-5. The $\bar{r}_c$s are calculated as weighted averages

$$\bar{r}_c(A;\mathbb{S}) = \frac{\sum_{r\in\mathbb{S}} c(r)\, r}{\sum_{r\in\mathbb{S}} c(r)}, \tag{2.1}$$

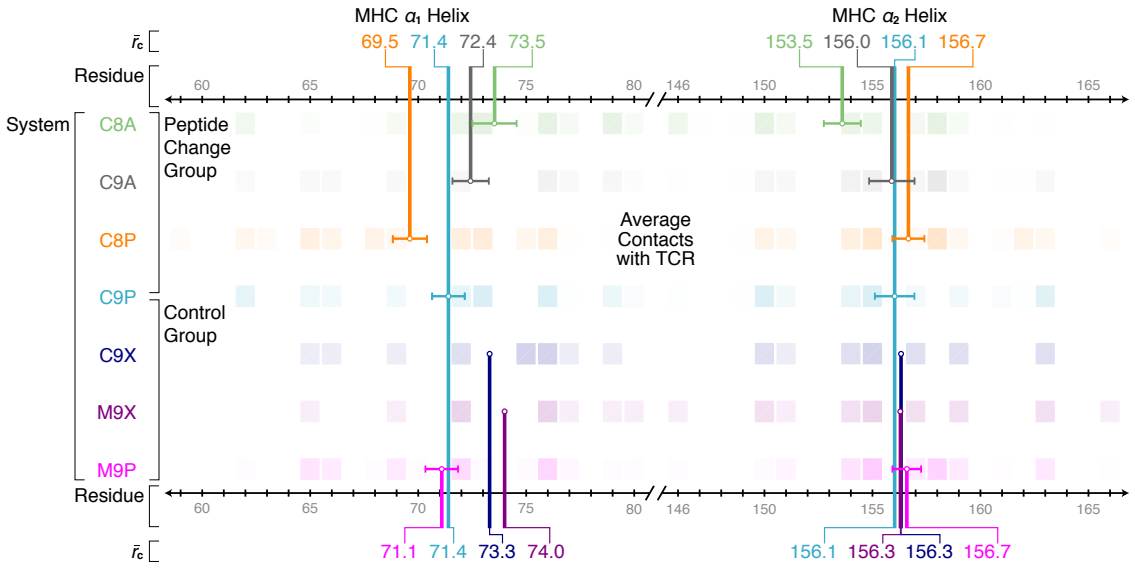


Figure 2-5: The changes in TCR/MHC contacts upon peptide mutation from C8P to C8A are numerically significant. The values of $\bar{r}_c$ for each system are plotted above on the same axes for comparison. $\bar{r}_c$ represents the "center of mass" of the contact distributions, and the difference induced in $\bar{r}_c$ by peptide mutation from C8P to C8A is clearly bigger than the difference between any systems in the Control Group. In between the $\bar{r}_c$ axes, a color-coded picture of the contact distributions in Figure 2-3 and Figure 2-4 is shown, in order to provide a qualitative picture of "contact mass": the more brightly colored a particular square is, the more contacts with TCR it makes. Errors signified by error bars emanating from the circular data points are in s.e.m., except for the crystallographic C9X and M9X systems.

where $\mathbb{S} \in \{\alpha_1, \alpha_2, \text{pep}\}$ is a label for a pMHC segment (either the α_1 helix, the α_2 helix, or the peptide) of a TCR/pMHC system A, r is the sequence number of a residue in $\mathbb{S}$ (denoted along the ordinate), and $c(r)$ is the average number of contacts that residue r makes with TCR (the length of the bar at position r that corresponds to system A in Figure 2-3 and Figure 2-4). Thus, $\bar{r}_c(A;\mathbb{S})$ represents the average position in the sequence where the TCR contact distribution for pMHC segment $\mathbb{S}$ is centered in system A, i.e. the contact center of mass. Differences between the means of two distributions representing different systems reflect underlying differences between the TCR/pMHC contact distributions of those systems (Figure 2-5), and we will use the notation $\Delta\bar{r}_c(A_1 - A_2;\mathbb{S}) \triangleq |\bar{r}_c(A_1;\mathbb{S}) - \bar{r}_c(A_2;\mathbb{S})|$ to refer to these differences throughout the paper.

2.4.5 Mutations to the CDR3$_\alpha$ loop of the TCR do not produce significant changes to the footprint

To verify the predictive quality of the simulations, we first examined how well they reproduce the experimental result that CDR3 mutations do not significantly alter the footprint [82]. Figure 2-3 and Figure 2-5 compare the contact distributions and $\bar{r}_c$s of C9P to M9P, and C9X to M9X—an intra-method, inter-TCR comparison—to assess the effect of the TCR mutation (2C → m6) on TCR/MHC contacts. The contact distributions of the crystallographic C9X and M9X systems are very similar to one another, except for a few stray contacts in M9X near the ends of the helices and two contacts with Arg75 of the MHC, which are only present in the C9X structure (Figure 2-3). Like their crystallographic counterparts, M9P and C9P are also similar to one another, as reflected by the small difference in their $\bar{r}_c$s (Figure 2-5). However, M9P has approximately one more average TCR contact than C9P in several places where MHC/TCR contacts already exist in C9P. The additional average contacts are distributed almost evenly throughout the entire contact interface. Due to the nor-

malization applied in Eq. 2.1, this difference is not reflected in the $\bar{r}_c$(M9P;$\{\alpha_1, \alpha_2\}$) values, and therefore, no significant shift is observed in the location of the footprint on the MHC. The even distribution of additional contacts suggests that m6 TCR interacts strongly with the pMHC. This is consistent with the experimental finding that the m6 TCR has a higher affinity for the L^d-QL9 ligand [82] and with the computed binding free energies reported in Supplemental Table A.1.

As a further test of quality, we compare the QL9 simulations (C9P and M9P) to their corresponding crystal structures (C9X and M9X)—the only direct experimental comparison available—in the bottom of Figure 2-5. This inter-method, intra-TCR comparison yields larger differences in $\bar{r}_c$ between simulation and experiment than the previous comparison, especially for the α_1 helix. The differences may reflect the large numbers of crystal contacts (476 crystal contacts for C9X and 874 for M9X), that are present only in the crystal structures and are absent in our simulations. Further analysis shows that, in the α_1 helix, 74% of $\Delta\bar{r}_c$(C9P −C9X;α_1) is due to the two TCR contacts with residue Arg62 on the α_1 helix of the MHC which develop in the C9P simulation, but are absent in the C9X structure. Other contact patterns are mostly conserved between simulation and experiment. Since such a large part of the difference can be accounted for by only one new contact location, the difference in $\bar{r}_c$s does not reflect the widespread shifts in pMHC contact locations that would accompany a significant change in the footprint. Although the inter-method $\Delta\bar{r}_c$(M9X −M9P;α_1) is larger than $\Delta\bar{r}_c$(C9X −C9P;α_1), and both are relatively large, they are still smaller than the largest $\Delta\bar{r}_c$ obtained from peptide mutation and therefore do not affect qualitative conclusions about its potential effects.

The smaller size of both of the intra-method, inter-TCR comparison $\Delta\bar{r}_c$s, relative to the inter-method, intra-TCR $\Delta\bar{r}_c$s, suggests that the simulations and crystal structures might have systematic, localized shifts from one another. Such system-

atic shifts may be caused, for example, by the absence of crystal contacts in the simulations. The two comparisons, taken together, suggest that despite some localized differences with the crystal structures, our simulations are consistent with the experimental result that mutation of the 2C CDR3$_\alpha$ loop does not induce significant perturbations in TCR/MHC contacts with QL9-L^d.

2.4.6 Mutations to the shortened antigenic peptide produce noticeable, local changes to TCR/MHC contacts

Among all the peptide mutant systems tested, two (C8P and C8A) show the largest shifts in $\bar{r}_c$ values, larger than any other pair of systems with the same TCR by at least a factor of two (Figure 2-5). The exact $\bar{r}_c$ values that underlie these differences are shown in Supplemental Table A.2.

Unlike the case of TCR mutation, the simple creation or annihilation of a single MHC/TCR contact cannot account for the large shift in $\bar{r}_c$. Rather, the shift is caused by the rearrangement of several contacts and an overall shift in the contact distributions of C8A and C8P to opposite ends of the sequence (relative to the wild-type C9P). This can be seen clearly in Figure 2-2b, where the CDR1 and CDR3 loops for both V_α and V_β have noticeably different average conformations in C8A than they do in C8P.

We also determined how widely distributed along the MHC the differences between C8P and C8A are (Figure 2-5). When comparing two footprints, we define a PMHC site as "changed" if it has 1.0 or more average contacts in one footprint, and 0.5 average contacts or less in the other. This choice of cutoff values is reasonable, as a PMHC residue with less than 0.5 average TCR contacts makes no contact with TCR more than half the time, but a PMHC residue with 1.0 average TCR contacts is in contact with the TCR on average once every time step in the trajectory. Using the above definition of change to compare C8A to C8P yields 8 changed sites out of

22, while comparing C9P to C9X or C9P to M9P yields only 4 out of 21 sites. The results are not very sensitive to changes of the cutoff values: if either parameter were varied by up to 10% in either direction, only two changed sites would lose their label, and in both cases, it is because the high value is very close to 1.0, and would cease to qualify as a contact if the upper threshold were raised to 1.1. The comparatively large number of changed sites demonstrates that differences between the C8A and C8P footprint are widely distributed and represent changes throughout the footprint.

Figure 2-5 also shows that the wild-type C9P contact distributions are centered between those of the C8P and C8A systems for both MHC helices. Thus, relative to the original, wild-type footprint, the lone deletion (C8P) and deletion-with-mutation (C8A) have opposite, non-additive effects, possibly due to how greatly deletion can affect the space available to the peptide. Taken as a whole, these effects are relatively localized to the TCR-MHC interface, and they do not involve the kind of structural rearrangement of TCR/MHC orientation observed when comparing the allo 2C/QL9-L^d(C9P) to 2C/SIYR-K^b (CSK) (Figure 2-2c), nor do they involve large rearrangement of the CDR loops themselves. This perspective is consistent with assessments of TCR/pMHC interfaces that highlight the relative inflexibility of the CDR loops [90, 91]. Nevertheless, it is clear that mutation of the antigenic peptide, unlike mutation of the CDR3$_\alpha$, can induce noticeable rearrangement of local contacts in the footprint.

Most importantly, these peptide mutation results are completely consistent with the findings of Felix et. al [92], which implied that changing the peptide can make changes to contacts between MHC and TCR throughout the footprint. They are also consistent with a previous crystallographic study of dissimilar peptides in the same TCR/allo-MHC complex [19]. Performing $\bar{r}_c$ calculations on the crystal structures compared in that study showed that the contact distributions of the dif-

ferent peptide mutants differed by as much as two residue positions—much more than C9X differed from M9X (Supplemental Table A.2).

Recent crystal structures obtained by MacDonald et. al [107] show the LC13 TCR in complex with different peptides in the same allo-MHC. In both cases, the TCR produces a remarkably similar footprint on the substrate. In this situation, however, no residues were removed from the peptide, so the physical impact of mutation was not as severe, despite several sequence differences between the two peptides. Such results indicate that not all kinds of peptide mutation are equivalent: while allopeptide mutations can have an impact on the TCR/pMHC footprint, this is not necessarily the case for a particular TCR/pMHC pair and a particular mutation—in the end, the specific interplay of TCR, peptide, and MHC interactions determines the ultimate interface structure.

2.4.7 Peptide mutations can impact the topology of the interface by changing the ratio of TCR V_α and V_β chain contacts

A particular feature of the 2C contact distributions is that the ranking of $\bar{r}_c$ values for the α_1 helix is the inverse of the ranking for the α_2 helix. That is, $\bar{r}_c(\text{C8A};\alpha_1) > \bar{r}_c(\text{C8P};\alpha_1)$, but $\bar{r}_c(\text{C8A};\alpha_2) < \bar{r}_c(\text{C8P};\alpha_2)$ (Figure 2-5). Since the sequence of the MHC α helices is organized in such a way that the N-terminus of one helix is close in space to the C-terminus of the other helix and *vice versa* (Figure 2-1), what appears as shifts in opposite directions along the sequence actually corresponds to a shift of the contact footprints in the same direction in space. Specifically, the shifts indicate that for C8P, more TCR contacts occur on the part of the MHC near the N-terminus of the peptide (the left end in Figure 2-1b), while for C8A, more TCR contacts occur on the part of the MHC near the C-terminus of the peptide (the right end in Figure 2-1b). The more space available to 8-mer peptides may allow this "rocking" flexibility, as is discussed in the following section.

Table 2.3: Peptide mutation can affect the ratio of V_β/V_α contacts with pMHC.

System	Avg. V_α/pMHC cont.	Avg. V_β/pMHC cont.	Ratio V_β/V_α cont.[a]
C9P	23.57 ± 0.06	16.91 ± 0.05	0.71 ± 0.08
C9A	23.20 ± 0.05	15.89 ± 0.07	0.68 ± 0.09
C8P	26.54 ± 0.06	21.17 ± 0.05	0.79 ± 0.08
C8A	15.41 ± 0.06	22.55 ± 0.06	1.46 ± 0.08
M9P	29.85 ± 0.05	19.03 ± 0.07	0.63 ± 0.09
C9X	25	23	0.92
M9X	27	22	0.81

[a] In C8A, TCR contacts with pMHC are biased toward the V_α chain, while in C8P, they are biased toward the V_β chain. The difference between average V_α contacts between C8A and C8P is larger than any of the differences obtained from comparing corresponding simulation and crystal structure data, or from comparing C9P to M9P. Errors reported are s.e.m.

From the diagram of the footprint in Figure 2-1b, it is clear that the V_α chain of the TCR is oriented closer to the C-terminal end of the α_2 MHC helix (on the right in the figure), and the V_β chain of the TCR is oriented closer to the C-terminal end of the α_1 helix (on the left in the figure). Combining this structural knowledge with the contact data discussed above, we expect that C8P would contain more MHC contacts with V_α, and C8A would contain more MHC-V_β contacts. In other words, the change from C8P to C8A would increase the ratio of V_β to V_α contacts. Table 2.3 shows that this is the case. In C8P, L^d makes 27 contacts with V_α and 21 contacts with V_β, while in C8A, L^d makes only 15 contacts with V_α and 23 contacts with V_β. The changes result in a shift of the V_β/V_α ratio from 0.79 to 1.46 as C8P changes to C8A. Figure 2-2 provides a structural explanation of the drop in V_α contacts, showing that in C8A, the CDR1_α loop tends away from the MHC, unlike in C8P.

These results demonstrate how the mutation of the fifth position of the shortened peptide from a proline to an alanine (C8P $\rightarrow$ C8A) can cause rearrangement of the contact topology, affecting not only which MHC residues come into contact with TCR, but also which TCR residues contact the MHC.

2.4.8 Mutations to the 9-mer peptide produce less contact rearrangement than mutations to the 8-mer peptide

By looking at all four peptide variants together, we can compare the relative impact of the mutation of the Ala6/5 residue on the footprint in both the 9-mer and 8-mer peptides. When we compare $\Delta\bar{r}_c(\text{C8A} - \text{C8P}; \{\alpha_1, \alpha_2\})$ to $\Delta\bar{r}_c(\text{C9A} - \text{C9P}; \{\alpha_1, \alpha_2\})$, it is clear that the footprint is significantly more sensitive to mutation at the Ala5 position of the 8-mer peptide than it is to mutation at the corresponding Ala6 position of the 9-mer peptide. While mutation of the 9-mer peptide does change contacts throughout the footprint, these changes do not affect the values of $\bar{r}_c(\text{C9P}; \{\alpha_1, \alpha_2\})$ or $\bar{r}_c(\text{C9A}; \{\alpha_1, \alpha_2\})$ more than inherent differences between crystal structure and simulation. Thus, although they may be due to underlying differences in the contact interface, these differences are not large enough to be seen as significant.

The shorter peptide evidently allows more freedom for the α helices—and thereby the entire footprint—to rearrange in response to a mutation in the center of the peptide. Despite such rearrangement, the mutation seems to have a similar effect on TCR binding in both the 8-mer and 9-mer peptide systems, lowering the TCR/pMHC affinity by two orders of magnitude (Table 2.1). This lack of correlation between binding affinity changes and structural rearrangement may seem counterintuitive, but it is in fact consistent with other results that describe sets of nearly identical TCR/pMHC crystal structures with highly varying TCR binding affinities [95].

The structural distinction that the footprint makes between mutations to the shortened and full-length peptide suggests that TCR and allo-pMHC have a complicated interaction landscape at their interface and therefore can respond to subtle changes in unique ways.

2.5 Discussion

The intent of our study has been to provide insight into the structural impact of peptide mutations on the alloreactive TCR/pMHC interface by comparing atomistic models of systems that have not yet been crystallized. In our simulations, mutation of the TCR's CDR3$_\alpha$ loop did not induce significant rearrangement of the TCR/MHC contacts, but certain mutations of the peptide made noticeable changes to local contacts throughout the TCR/pMHC interface.

Our results contribute to a large body of work aimed at elucidating the nature of the 2C TCR alloreactive response in particular and alloreactive responses in general. That work has yielded two differing pictures of the peptide's structural influence on the TCR/pMHC footprint: a peptide-centric model [84, 92, 93, 98], in which the peptide plays an important role in determining the footprint, and an MHC-centric model [82, 95, 104, 105], in which the footprint is determined by the MHC, with the peptide having virtually no effect. These views are sometimes simplified as a "tail wagging the dog/dog wagging the tail" debate, with the dog being the MHC and the tail being the peptide. Some experimental results, such as the MHC mutation experiments of Felix et al. [92] and the alanine-substitution work of Conolly et al. [84, 93], have been cited as support for the peptide-centric idea. Other experiments, specifically the work of Colf et al. [82], have been explained as support for the MHC-centric model.

Our findings serve as a synthesis and explanation of these apparently conflicting views. We show that, while CDR3 mutations do not necessarily influence the alloreactive contact footprint, peptide mutation, by contrast, can have an influence. The difference between the observed effects of these two kinds of mutations highlights the difference between influencing the TCR/peptide interactions by changing only the TCR's CDR3 residues and mutating the peptide itself, which is a more direct way to test the peptide-centric model. Our simulations reveal that

the two experimental approaches are not equivalent, as mutations to the peptide can directly impact the MHC helices and thus affect the chemical surface to which the TCR binds in ways that mutations of the CDR3 loops cannot. However, even in the cases where we found that peptide mutations broadly affected the footprint, these effects were relatively localized compared to the difference between self- and allo-MHC footprints made by the 2C TCR [82].

2.6 Conclusions

We thus predict that both MHC and peptide have a direct impact in determining the TCR/allo-pMHC footprint, with the peptide being able to "edit" the specific contacts after initial contact between the TCR and pMHC. One mechanism for this multilevel impact may involve the TCR "reading" the pMHC ligand in search of enough sufficiently favorable interactions. If enough interactions exist and the TCR/pMHC complex survives long enough, modest, specific rearrangements of the interface residues would occur to maximize the affinity [70, 91]. This mechanism is consistent with our findings and is also consistent with diverse other findings, including those of MacDonald et al. [107], because rearrangements after the initial scanning induced by the peptide could produce a similar final structure in initially different pMHC landscapes.

Looking forward, we believe that free energy simulations of TCR/allo-pMHC similar to those previously performed for TCR/self-pMHC [28–30] would be informative, providing more detailed free energetic data, such as a systematic comparison of the relative contributions of peptide mutation and MHC changes on the TCR binding affinity. Most importantly, testing the predictions made by our simulations by obtaining crystal structures of the C8P and C8A systems (or some analogues) and performing analysis similar to our own with these structures will shed more light

on the interplay between peptide and MHC on alloreactivity. We hope that our study will motivate such experiments.

Chapter 3

Dimensionality reduction techniques and visualizations for phenotype analyses of adoptive T-cell transfer melanoma therapy

3.1 Summary

Adoptive cell transfer therapy holds promise in cancer treatment. One of the key issues inhibiting its effectiveness is the ready identification of specific phenotypes responsible for anti-cancer activity. New multicolor FACS technology allows identifying cell types with many more surface markers, but this increases the dimensionality of the space and makes phenotype distributions more difficult to visualize. To enable visualization and analysis of the new, higher dimensional datasets, we applied techniques from multivariate analysis and developed unique visualization techniques to present the multidimensional data in a comprehensible format. This thesis chapter outlines and explains these methods and their application to two specific datasets: a *in vivo* set of T cells subject to different stimulation proto-

cols, and a *in vivo* set of ten patients subject to ACT. We found that our methods could redisplay the data in useful ways to identify several global trends, pick out specifically relevant phenotypes, and detect outliers. We hope that our insights can direct future stages of experiments to understand and improve this promising medical technology.

3.2 Introduction

Adoptive cell transfer (ACT) therapy is a promising technique for treatment of metastatic melanoma [124]. During ACT, cancer-specific T cells are activated and expanded *ex vivo*. The patient's own T cells are depleted, and the activated cells are transfered as a replacement. Many approaches exist for generating and activating the cancer-specific T cells before infusion. Regardless of the approach, the phenotype distribution of the transfer population before and after transfer strongly affects the result of the treatment [10]. Understanding which phenotypes lead to improved patient outcomes is therefore essential to the growth and development of ACT.

T cell phenotypes are most commonly measured with FACS immunotyping of the surface markers expressed on the surface of a population of cells. Early immunotyping implementations allowed simultaneous detection of only one or two cell surface markers. The restricted number of simultaneous measurements resulted in datasets that were easy to visualize directly in $1-3$ dimensions. However, this data could not measure higher-order correlations between surface marker levels.

Recent advances in multicolor FACS immunotyping have enabled the simultaneous detection of five or more surface markers on the same cell [33]. The higher-dimensional data provide a much richer description of the phenotypes present in

a cell population, but they are difficult to visualize and interpret. Attempts to visualize the data directly, such as color-coded scatter plots [125], suffer from a lack of clarity: even if all five phenotypes are represented with some combination of position and color, direct visual comparison of five-dimensional sets of data is elusive [126]. The recent work of Nolan and coworkers [127] provides a powerful technique for handling wide-spread variability in size and granularity of cells in FACS measurements and allows side-by-side comparison of high-throughput results from many conditions, but it does not address the question of interpreting the resulting multi-dimensional space of surface markers and phenotypes.

Given the difficulty of direct visualization, understanding the data requires analyses that extract or highlight relevant qualitative features. Dimensionality reduction and correlation analyses are powerful techniques for addressing specifically these types of problems, especially when combined with appropriate visualizations. They identify qualitative features in the data through calculation of similarity measures and projection of the data into revealing subspaces of lower dimension.

This thesis chapter describes the application of statistical and heuristic techniques for dimensionality reduction, comparison, and visualization of complex, high-dimensional FACS data of ACT cells obtained by our collaborators at UCLA [128, 129]. The data come from two datasets: a collection of *in vitro* phenotype measurements made on engineered T cells subjected to four activation protocols, and a collection of *in vivo* measurements made on a separate set of T cell populations adoptively transfered to a cohort of ten human patients.

3.3 *In vitro* data

The biological motivation and experimental details behind the collection of the *in vitro* dataset are fully described in the work of Tumeh, *et al.* [128]. Thus, the emphasis of this work will be on the statistical modeling, and only the essential biological detail will be described.

3.3.1 Methods

Experimental setup

Multicolor FACS immunotyping was performed on CD4 and CD8 T cells subjected to four different clinical-grade activation protocols. The four protocols resulted from two choices of activation method (anti-CD3 antibody OKT3 or anti-CD2/3/28 beads) and two choices of cytokine milieu (IL-2 alone or IL-2 with IL-15). These activation protocols, along with their abbreviations, are described in Table 3.1.

For each protocol, the cells were measured at five time points: 0, 2, 7, 15, and 21 days after activation. At every time point, three FACS experiments were performed on the population. Each experiment measured a separate set of surface markers, as described in Table 3.2. The five surface markers in each set define $2^5 = 32$ phenotypes, which come from all possible combinations of a cell having high (+) or low (−) expression of that particular surface marker. Thus, an example phenotype from group A would be CD25+ CD127− CD45RO+ CD44+ HLA-DR−. Every phenotype in a group is associated with a single data point for every time point and activation protocol. This data point is a cell frequency, a fractional number between 0 and 1 representing how many of the cells on average express this phenotype. For a given activation protocol and time point, the frequencies in each group necessarily sum to 1.

In summary, the dataset consists of 1920 data points, as described by the equa-

Cytokine	Activation	
	OKT3 antibody	anti-CD2/3/28 beads
IL-2	P_{00}	P_{01}
IL-2 / IL-15	P_{10}	P_{11}

Table 3.1: Engineered T cells were subjected to four activation protocols, resulting from two choices of activation method and cytokine milieu. The label given to each protocol is shown in the table.

Group	Surface Markers				
A	CD25	CD44	CD45RO	CD127	HLA-DR
B	CCR5	CCR7	CD45RA	CD137	PD1
C	CD27	CD28	CD57	CD62L	CD95

Table 3.2: Each cell population was measured for three separate groups of five surface markers each. Note that it is not possible to measure joint expression of surface markers in different groups directly.

tion below:

$$4 \text{ protocols} \times 5 \frac{\text{time points}}{\text{protocol}} \times 3 \frac{\text{marker groups}}{\text{time point}} \times 32 \frac{\text{phenotypes}}{\text{marker group}} \times 1 \frac{\text{data point}}{\text{phenotype}} = 1920 \text{ data points.} \quad (3.1)$$

Statistical analyses

The immunotyping experiments result in a wealth of data: too much to be readily visualized directly. Dimensionality reduction and analysis are needed in order to organize the data in a meaningful way and address important biological questions. The analyses performed will help address two key areas: how different are the effects of the four activation protocols on the phenotype distributions, and which phenotypes and surface markers are the most highly variable? The more variable a surface marker is, the more likely it is to represent an underlying biological process.

Both of these questions can be addressed with correlation and principal component analysis. These two related techniques describe and analyze which variables in a dataset are changing the most, how changes in one variable tend to impact another variable, and which subsets of variables tend to change in similar (or opposing) ways. The techniques can be effectively described without the details of the mathematics, which can be found in a broad spectrum of mathematical literature [38].

Fundamental to both of the analysis techniques is the notion of "variables" and "samples." Variables refer to a quantity that can change, such as the fraction of cells that express a surface marker, or the concentration of a compound. Samples refer to explicit observations of a variable. Datasets are often organized in tables, with each column labeled by a variable and each row labeled by a sample. This abstract assignment of variables and samples makes clear that the two notions are labels and can be assigned in multiple ways. One could imagine turning the data table on its side, flipping the assignment of the two terms. Such a transformation may not always have meaning, but it is often worthwhile and allows a different perspective on the data.

In the FACS data, the straightforward approach is to label the phenotypes as variables. The samples of the phenotype variables are then the frequencies of those phenotypes measured under all twenty treatment and time point conditions. The opposite label assignment is less intuitive but equally valuable. In this assignment, the variables are the twenty measurement conditions, each specified by an activation treatment and a measurement time. The samples of the condition variable are the frequencies of all 96 phenotypes measured under that condition. When both label assignments are used with the same analysis technique, an application of that technique is referred to for clarity as either a *phenotype* analysis (the variables are phenotypes, the samples are experimental conditions) or a *condition* analysis (the

variables are experimental conditions, the samples are phenotype frequencies).

For either assignment of variables and samples, there is a quantity—the *covariance*—that describes how one variable fluctuates with respect to another. If, on average, variable A tends to go up as variable B goes up, their covariance will be positive. If the opposite happens, the covariance will be negative. In the calculation, the samples represent the specific values of A and B used to determine this quantity. If $A = B$, the covariance is the same as the variance of A.

Two examples of covariance calculations illustrate the impact of the assignment of variables and samples. In an example phenotype covariance,

$$A = \text{CD25+CD127−CD45RO+CD44+HLA-DR−, and}$$

$$B = \text{CD25−CD127+CD45RO+CD44+HLA-DR−.}$$

The frequencies of both phenotypes are compared at the same time points for all four treatment conditions (the samples), and based on their similarity, a single covariance number is calculated that relates the two phenotype variables. Conversely, in a sample condition covariance, A = day 2/IL-2/OKT3, and B = day 7/IL-2,IL-15/OKT3. In this formulation, the samples are the individual phenotype frequencies for each of those conditions. That is, instead of comparing the levels of two phenotypes under all measurement conditions, the comparison is now between two measurement conditions, over *all* phenotypes. This correlation can be interpreted as a measure of the similarity of the phenotype frequency distributions for the two conditions.

The larger the values of A or B are, the larger the magnitude of their covariance will be, regardless of its sign. This is not always desired, for example, when quantities are measured in different units. Another example comes in the case of comparing measurement conditions above. When comparing phenotype distribu-

tions across measurement conditions, there is no reason to give conditions with more variable distributions a higher covariance with other distributions: all that is of concern is the way that two conditions' distributions compare relative to one another. In such cases, the covariance can be rescaled to be a number between -1 and 1. The rescaled covariance is known as the *correlation*, and has similar properties to the covariance except its dependence on magnitude. For a set of n variables, n^2 covariance or correlation values can be arranged into a matrix whose entry at row i, column j represents the covariance between the ith and jth variables. This matrix is symmetric along its diagonal, the entries of which are either the variances of the variables (in the case of covariance) or all ones (in the case of correlation). Two examples of a correlation matrix are shown in Figure 3-1.

Visual inspection of the correlation and covariance matrices in Figure 3-1 can already yield many qualitative insights into the structure of the data. For example, for CD4 cells, the bottom-right corner of of the correlation matrix shows that the phenotype distributions collected on day 21 are much more similar to themselves than to any of the other distributions, regardless of treatment. Furthermore, the data from days 2 – 15 are relatively similar to one another, and day 0 is unique. We can immediately observe that experimental protocol plays a secondary role to time in describing the collected phenotype distributions.

But correlation matrices may not always exhibit such visually distinct correlated and uncorrelated groups. We must therefore extract indicators from the data that describe the important qualitative aspects of the correlations in the data. This extraction process is known as Principal Component Analysis (PCA), and the extracted indicators are known as principal components (PCs) [38]. Principal components are weighted sums of the original (manifest) variables defined by the sets of weights (loadings) used in the sum. They are defined to be entirely uncorrelated with one another, while simultaneously capturing as much of the variance in the

original data as possible. When used to analyze a covariance matrix, they represent correlated clusters of the most dynamic and widely changing variables; when used to analyze correlation matrices, they represent clusters of highly correlated variables. There are as many components as original variables, but, because each component is designed to capture as much of the data's variances as possible, only the first handful tend to matter. These first few components define a subspace onto which the data can be projected, similar to cutting through a cloud of points with an aptly placed plane. The projection can reveal important grouping in the data that is not visible by examining it directly. The technical details of the PCA are described in Chapter 1.

Finally, because the phenotypes are composed of surface markers, the total expression frequency of a surface marker is the sum of the expression frequencies of all the phenotypes that express it. For example, if $f(x)$ represents the total expression frequency of x at a particular measurement condition, then the following equation holds:

$$
\begin{aligned}
f(\text{CD25+}) = \\
& f(\text{CD25+CD127}-\text{CD45RO}-\text{CD44}-\text{HLA-DR}-) \quad + \\
& f(\text{CD25+CD127}-\text{CD45RO}-\text{CD44}-\text{HLA-DR+}) \quad + \\
& f(\text{CD25+CD127}-\text{CD45RO}-\text{CD44+HLA-DR}-) \quad + \\
& f(\text{CD25+CD127}-\text{CD45RO}-\text{CD44+HLA-DR+}) \quad + \\
& \qquad \cdots \\
& f(\text{CD25+CD127+CD45RO+CD44+HLA-DR+})
\end{aligned}
\tag{3.2}
$$

Also, because all surface markers are exclusively either high (+) or low (−), it is

clear that

$$\begin{aligned} f(\text{CD25}-) &= 1 - f(\text{CD25}+) \\ f(\text{CD25}+) &= 1 - f(\text{CD25}-). \end{aligned} \tag{3.3}$$

By calculating and analyzing the expression frequencies of surface markers and not only individual phenotypes, we can see if and how specific phenotype fluctuations fit into a larger picture of surface marker fluctuations. For example, the CD25+ CD127− CD45RO− CD44− HLA-DR− phenotype may be highly expressed, with a frequency of 0.7. We would like to know which surface markers are associated with that high frequency. In order to do this, we calculate the total expression frequencies of each of the surface markers in the phenotype. Since all the frequencies are nonnegative, these total expression frequencies will be larger than 0.7. However, the closer a surface marker's frequency is to 0.7, the larger the percentage of that frequency comes from CD25+ CD127− CD45RO− CD44− HLA-DR−. If, for example, $f(\text{CD25}+) = 0.72$, and the other surface markers

$$f(\text{CD44}-) = f(\text{CD45RO}-) = f(\text{CD127}-) = f(\text{HLA-DR}-) = 0.99,$$

it is clear that the CD25+ surface marker is the only informative part of the phenotype: since almost all of the cells in the sample are CD44−CD45RO−CD127−HLA-DR−, it does not matter that this highly expressed phenotype contains those surface markers. But regarding CD25, it is clear that the CD25+ CD127− CD45RO− CD44− HLA-DR− phenotype makes up over 97% of all CD25+ cells (as opposed to only 70% of the other surface markers). Based on such interplay between the surface marker levels and the individual phenotypes, we can analyze general trends, such as an increase in CD25+ cells, and also identify the specific phenotypes that are responsible for those trends, such as CD25+ CD127− CD45RO− CD44− HLA-DR−.

3.3.2 Results

The analysis techniques described in the Methods section were combined to extract systematic trends in the data and present it visually in a way that highlighted the most relevant changes.

Measurement condition correlations

In order to assess the effects of activation protocol and time on phenotype distributions, correlations between measurement conditions were calculated and are displayed in Figure 3-1. For this analysis, all three groups of phenotypes (96 in total) were used as samples to compare the phenotype distributions of every pair of experimental conditions.

The largest trend in the data is immediately visible: clustering of correlations around days. The indication is that, for CD4 cells, the distributions at Day 0 are unique. Then, on days 2, 7, and 15, the distributions are all relatively similar. Distributions for the same day are usually more similar to one another than other days, and activation protocol does not appear to play a large role in distinguishing the experimental conditions for that day from one another. Finally, by day 21, the population has shifted to yet another, dissimilar phenotype distribution, which is almost invariant to the choice of activation protocol. CD8 cells exhibit a similar trend, except that the day 2 conditions do not belong to the same "block" to which the day 7 and day 15 conditions belong. In summary, while the activation protocol may impact other factors, such as population size, growth rate, and CD4/CD8 balance, it is clear that, within the separate CD4 and CD8 populations, the main factor influencing the distribution of phenotypes is time, not activation protocol.

The second most visible conclusion is the observance of an "outlier" in the IL-2/OKT3 protocol at day 7. This distribution is much more similar to itself than to any other experimental condition for CD8 cells, and expresses a similar uniqueness

for CD4 cells. This unexpected violation of the trends established by all of the other measurement conditions leads to the conclusion that some experimental circumstances affected those cells in an unexpected way. Our conclusion, arrived at with no knowledge of the specific experimental conditions, was later supported by the experimentalists, which displays the method's ability to successfully predict relevant properties from only the data.

Thus, by simply describing the correlations between phenotype distributions over the multiple measurement conditions, it is possible to observe global trends in the data and identify outliers. Now that the general trends have been identified, the specifics of which phenotypes are highly expressed in which of the experimental conditions can be determined by direct examination of the phenotype distributions for the conditions of interest, or with another visualization method, discussed below.

Time evolution of phenotype distribution

In addition to observing global trends, it is possible to highlight specific fluctuations in phenotype levels. Unfortunately, to display all of this data together would be overwhelming. We therefore developed a technique to display what the most important phenotypes are, given a particular FACS experiment, and show which surface markers are most closely associated with their expression. To display this data, we have developed a new visualization technique, the "subway plot," shown in Figures 3-2 – 3-7.

The subway plot, so named for its resemblance to maps of subway routes in urban areas, shows a select subset of the phenotypes collected during a FACS experiment and subjected to an activation protocol. Since the first principal component in covariance PCA analyses is known to identify the most variable factors with large weights [38], the subset is determined by selecting the most variable and highly

expressed phenotypes, by selecting all phenotypes with large loadings in the first principal component of a phenotype covariance PCA. In this way, the multidimensional nature of the phenotypes has been reduced to a single dimension without sacrificing the complexity of correlations between multiple surface markers.

The subway plot shows which surface marker frequencies are influenced by a phenotype. For every time point, each phenotype in the subset is compared to every surface marker it expresses. Any surface markers which are sufficiently widely expressed in the entire population and whose expression comes mostly from that phenotype are identified with large circles as "significant." Circles from the same phenotype on neighboring timepoints are connected with thick lines for visual clarity. Specifically, for a surface marker to be identified at a particular time point, the phenotype must constitute more than 5% of the population and be responsible for more than 20% of all cells which express that surface marker. Furthermore, there can be no more than three such markers, and the contributions of those markers must be at least 150% of the average contributions of the other markers. The second set of criteria prevent a situation in which a phenotype makes significant contributions to too many surface marker levels. Because of the stringent criteria for significant surface marker identification, some time points are identified with no surface markers at all. When significant markers do exist, they are further identified by a pair of numbers ($p_1 \times p_2$): expression percentages. p_1 and p_2 show that the phenotype in question accounts for p_1% of all cells expressing (or not expressing) the surface marker, and that p_2% of all cells express this surface marker. The corresponding expression level of that phenotype at that time point is thus $p_1 \times p_2$. Because subway plots display three-dimensional data, it is often difficult to identify a point's position on the (x, y) plane. For this reason, the subway plot "grounds" each point with a line extending down to the (x, y) plane. The position where the line meets the (x, y) plane unambiguously identifies the x and y coordi-

nates of that point. The tick marks on the line correspond to the tick marks on the z axis at edges of the plot, unambiguously identifying its z coordinate.

The subway plots highlight the same general trends as displayed by the correlation analysis: that, on the whole, time has a much larger impact on the phenotype distribution than activation protocol. However, the subway plots also allow identification of specific surface markers that are commonly expressed in particular conditions. For example, in Figure 3-4, Group C clearly highlights the broad emergence of CD62L– cells in day 21. In all cases, approximately half of this population is accounted for by the CD27– CD28+ CD62L– CD57– CD95+ phenotype. Another example is available in Figure 3-6, Group B, where the subway plot shows the clear emergence of CD45RA– cells at the 7- and 15-day time points, fueled by the CCR5– CD45RA– CCR7– CD137– PD1– phenotype. By combining both general trends with specific data, the subway plots allow comprehensive visualization and comprehension of the key features in the phenotype distribution data.

3.4 *In vivo* data

Like the *in vitro* data, the goal in analyzing the *in vivo* data was to identify specific phenotypes with biologically relevant properties and to obtain a global perspective on significant trends in the data [129].

3.4.1 Methods

Experimental setup

In vivo phenotype distributions were obtained from a cohort of ten patients undergoing a trial ACT therapy. Genetically engineered CD4+ and CD8+ T cells were stimulated *ex vivo* and transfered to a cohort of ten patients. Phenotype distributions of T cells were collected from the patients' blood at time points between 0

Patient	Collection Time (days)							
F5-1	0	9	14	30			72	90
F5-2	0	9	15	30			76	
F5-3	0	9	15	29	45	57	71	86
F5-4	0	7	15	29	44			
F5-6	0	7	15	30	45	60	73	87
F5-7	0	7	15	30	45	59	76	89
F5-8	0	7	15	30	46	60		84
F5-9	0	9	14	30	43	59	78	
F5-10	0	7	15					
F5-11		7	18	32				

Table 3.3: Phenotype data was collected from each patient for up to nine time points between 0 and 90 days after transfer. Days were grouped according to a hierarchical clustering approach using their Euclidean distances. Data for some patients was collected beyond 90 days but is not used in the analysis.

Group	Surface Markers			
A	CD25	CD45RO	CD127	HLA-DR
B	CCR5	CCR7	CD45RA	PD1
C	CD27	CD28	CD62L	

Table 3.4: In the *in vivo* data, each patient's T cells were measured for three separate groups of three to four surface markers each. Note that it is not possible to measure joint expression of surface markers in different groups directly.

and 90 days after transfer. The multicolor FACS data generated from the *in vivo* trials was similar to the *in vitro* data but contained three substantial differences. First, instead of four activation protocols, the data now consisted of phenotype distributions from ten patients. Second, each patient had more time points and the time points were no longer the same for all patients, as shown in Table 3.3. Finally, the phenotypes themselves were different. They are composed of fewer surface markers and the groupings have changed. The *in vivo* surface markers are displayed in Table 3.4.

Statistical analyses

Given the success of statistical analyses to extract qualitative information and from the *in vitro* data, we extended the techniques to describe and analyze the *in vivo* data. The large number of patients (as compared to the relatively few *in vitro* activation protocols) meant that the subway plots developed to analyze the *in vivo* data were no longer coarse enough to identify global trends: to compare ten subway plots simultaneously is much more difficult than to compare four. Instead, other methods were employed to identify relevant trends.

Specifically, phenotypes which decayed or grew very rapidly and phenotypes which varied greatly were of biological interest. In addition to identifying particular phenotypes which exhibited interesting behavior, Principal Component Analysis was used to represent the entire dataset in a reduced-dimensional space which identified the phenotypes which corresponded most strongly with a particular patient or time point.

Finally, because of the many time points available for each patient, a regression model was created to establish a relationship between phenotypes and the time during which they are expressed. The regression model was developed with Partial Least Squares Projection onto Latent Subsets (PLS, also called Partial Least Squares Regression), a statistical technique that combines PCA with least-squares regression to produce correlated subsets of variables, a reduced-dimensional mapping, and a regression model that is resistant to over-fitting. The time-based PLS model, based on the work of Rivet and coworkers [10], provides a concrete indicator of how strongly each phenotype is associated with early or late time points. It can also be used to predict the age of a population of adoptively transfered T cells solely from its phenotype distribution. The remaining technical details of the PLS model are described in Chapter 1.

3.4.2 Results

Outliers and inconsistent data detection

To identify global trends between patients and time points, a condition correlation analysis was performed on the patient data. Its schema mimicked the *in vitro* correlation analysis: the individual patients and time points were treated as variables, and correlation coefficients were calculated across variation in phenotype expression levels. The results of the condition correlation analysis are displayed in Figure 3-8.

The structure of the correlations displayed in Figure 3-8 allows us to identify inconsistencies in all CD4+ T cell observations and CD8+ T cell observations from patient F5-6. The CD4+ correlations show considerable cross-patient variability, which is starkly different from the globally positive correlations seen in the CD8+ data. The variability is likely associated with the low CD4+ cell counts obtained from patients' blood and is therefore attributed to experimental error. Similarly, correlations in CD8+ phenotype distributions from patient F5-6 show a visible contrast with data collected from the other patients. This inconsistency is also associated with other indicators of experimental error in the data from patient F5-6. Because both the CD4+*a*nd F5-6 data were associated with other indicators of experimental error, they were discarded from further analysis.

Although the correlation analysis highlights experimental sources of error, its ability to identify inconsistencies without prior knowledge speaks to the power of the technique. In similar situations, it is clear that data displaying a unique correlation pattern should be scrutinized in order to ensure that the source of the unique pattern was not an experimental artifact.

Variability and dimensionality reduction

As with the *in vitro* data, it is important to identify which *in vivo* phenotypes vary the most and how they change across time and patient. For this reason, a phenotype covariance PCA and correlation PCA were performed on the data. The covariance PCA identified the most variable phenotypes and the correlation PCA projected the data into reduced dimensions that displayed patient-based clusters and identified which phenotypes discriminate best between patients. These techniques provide powerful insight into the most substantial sources of variation in the data.

By comparing the variances of phenotypes to the magnitude of their loadings in the first principal component of a covariance phenotype PCA, Figure 3-9 displays a clear separation of a small subset of highly varying phenotypes. Both the variance and the first PC loading are indicators of variability. They are nearly synonymous to some extent, as the $r > 0.9$ correlation coefficient indicates. However, they differ enough to allow discrimination in the choice of variables. This is because the first PC loading also takes into account covariance with other phenotypes to a certain extent. Thus, a phenotype which by itself does not have a particularly high variance but covaries strongly with other variables which do have large variances may still have a high loading in the first PC. By comparing variability with these two metrics, Figure 3-9 is able to identify six highly varying phenotypes. In particular, the CD27– CD28– CD62L– phenotype stands out as both the phenotype with the largest variance and the largest PC loading.

In addition to comparing variability, PCA can be used to identify the sources of variation in the data and project it onto a subspace that is defined by those sources. Figure 3-10 displays a biplot that summarizes a wealth of global information about the phenotype data. A biplot is a plot specifically designed to plot the results of a PCA together [39]. The content is composed of two parts: a scatterplot of PCA *scores* and radial lines representing the PCA *loadings*. The loadings, as described

before, are the weights that define a principal component in terms of the manifest variables (in this case, the phenotypes). They show how the plane defined by the two PCs is oriented in the space of the phenotypes, i.e, which phenotypes impact the PCs, and in what way. The scores are the representation of the observations in terms of the principal components. Each point on the score scatterplot corresponds to a single condition: a patient and timepoint. Rather than being identified by a whole distribution of phenotypes, it is now identified by only two points.

Plotting the scores and loadings together produces a visual interface greater than the sum of its parts. The score scatterplot shows a visual representation of the data in two dimensions. The simultaneous overlay of the loadings on top of the scores comes with an additional benefit. Because the x and y coordinates of the biplot refer to the same principal components, the coordinates of a loading and a score are related. If PC 2 is displayed along the x axis, a phenotype with a high loading in PC 2 will have a large x coordinate, and any condition with high levels of that phenotype will also have a large x coordinate. In this way, the geometric distance between a phenotype's loading point and an experimental condition's score point is a proxy for how highly expressed that phenotype is under those conditions. Finally, combining color and size to identify a scatterplot point with both its patient and relative sample time allows complete visual identification of all relevant information about the samples.

The largest trend revealed by the biplot is the difference between the phenotype distribution upon infusion and all other timepoints. This is represented by the entire first and fourth quadrants containing scatterplot points from "day 0" measurement only. The distinction between day 0 and the other timepoints is described almost exclusively by the first principal component, which explains the largest fraction of the total variation in the data. Thus, it is the largest trend in the data, larger than any other correlation between any of the variables. The loading

rays that point most strongly in the $+x$ direction, such as CD27+ CD28+ CD62L–, describe the phenotypes that most readily distinguish themselves by their appearance on day 0.

The second most significant trend is the patient-by-patient clustering of the data, clearly revealed through the projection onto the first two principal components. In this projection, not all patients cluster closely together, e.g. F5-3 and F5-7. But many of the other patients, e.g. F5-8, F5-9, and F5-1, clearly localize in their own area of principal component space. This means that, on a coarse level, after the phenotypes have transformed from their original distribution, measured upon infusion, they tend to converge to a patient-specific distribution that does not change much with time. Some of distributions are clearly associated with a particular phenotype, such as the high expression of CD27– CD28– CD62L– in F5-1. The patient by patient clustering shows that cross-patient variability is a continuing challenge to adoptive T cell therapies. Because individuals' immune systems vary so starkly, due to a host of genetic and environmental factors, it is difficult to know the specifics of what effects a particular T cell clone will have. With more patient data, these differences could be quantified and understood.

Time-based regression

In addition to understanding what the biggest sources of variation and correlation in the data are, it is also useful to identify which phenotypes grow and which ones decay. A PLS model that predicts the time from the phenotype distribution, based on the work of Rivet *et al.* [10], automatically separates out the strongest growing and decaying phenotypes with only a single tunable parameter. This model reaches its highest performance with only a single principal component. Phenotypes with positive loadings in this PC correspond to growing phenotypes, and negative loadings correspond to decaying phenotypes. Plots of these two groups

are shown in Figure 3-11.

The two groups exhibit clear growth and decay phenomena. It is interesting that such a simple model can capture much of the essential dynamics. Phenotypes were assigned one of the two groups selected if their magnitudes were greater than 50% of the maximum expression that day.

Despite the cross-patient variability, our ability to construct an accurate regression model of the time from the phenotype data suggests that there are consistent time-based phenomena occurring in the patient's response to adoptive T cell transfer. The time-based structure helps us understand in detail which phenotypes emerge later after transfer. By studying the late-emergent phenotypes in greater detail, we will seek to identify the underlying biological reasons for their emergence.

3.5 Discussion

In this thesis chapter, we have shown the power of correlation analysis, PCA, and PLS for analyzing newly available multidimensional FACS data. We have also shown how biplots can identify a wealth of global information about a dataset, and developed a visualization technique—the subway plot—for comparing specific time trajectories of phenotype distributions, while identifying the most relevant surface markers whose variation determines the phenotypes' fluctuations. These are the surface markers to which further experimental effort should be devoted.

Our analysis methods are powerful techniques that should be incorporated into the workflow of newer, more complicated data collection schemes. Nevertheless, these techniques are still subject to limitations. First, the linear aspect of the modeling may prove too simplistic for certain relationships, such as oscillatory variation. This can be mitigated to a certain extent through inverse transformation, if

the form of the data is known. For example, an exponential relationship can be linearized by taking the logarithm of the exponentially growing quantity. Furthermore, the simplicity of the PLS model does not take into account systematic noise in the phenotype expression levels that is uncorrelated with the age of the cells. Filtering this data out via methods such as O-PLS and its extensions [45,57,59,130] would be a useful next step.

Finally, experiments to test the insights revealed by the analyses are essential. Our analyses have provided a system for identifying where experiments should be directed, but they cannot bring additional biological meaning to the problem. For example, if our PLS model identifies CD27– CD28– CD62L– is identified as a particularly phenotype in determining T cell age, we must understand why this is so. What other surface markers are present on CD27– CD28– CD62L– cells? Are these markers also good indicators of late time? How do CD27– CD28– CD62L– cells function? Which immunogenic properties do CD27– CD28– CD62L– cells exhibit? Experimental work is currently being done in this direction, but has not yet produced results. With the upcoming experimental data, our analysis work has tremendous potential for expansion, and we hope the techniques developed in this thesis chapter will be used later on to guide experiments that reveal the behavior and theraputic abilities of adoptively transfered T cells.

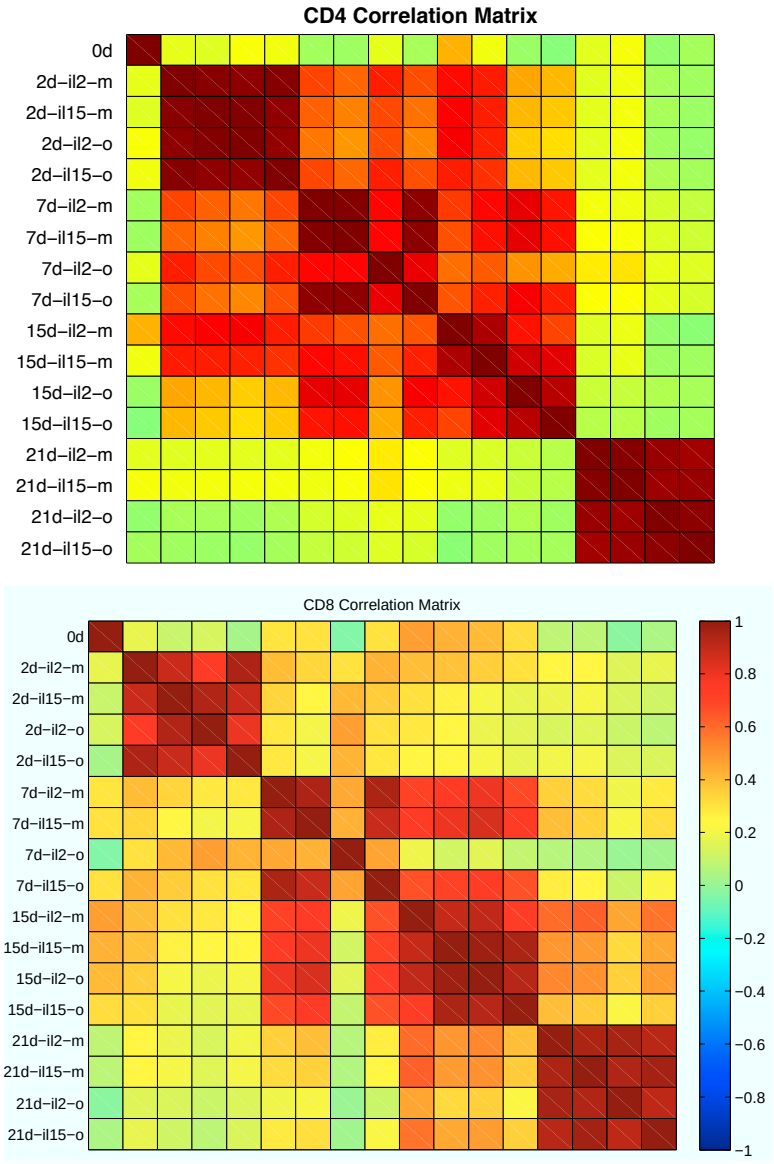


Figure 3-1: Measurement condition correlations highlight the primary impact of time on phenotype distribution. Labels for each column correspond to the row labels, moving from left to right. Each dash-separated section of the labels describes a particular experimental condition. The first section describes the day, 0 – 21. The second section describes the cytokine regimen, and the third section describes the activation protocol: "o" for OKT3 antibody, and "m" for andibody-coated beads.

Group A

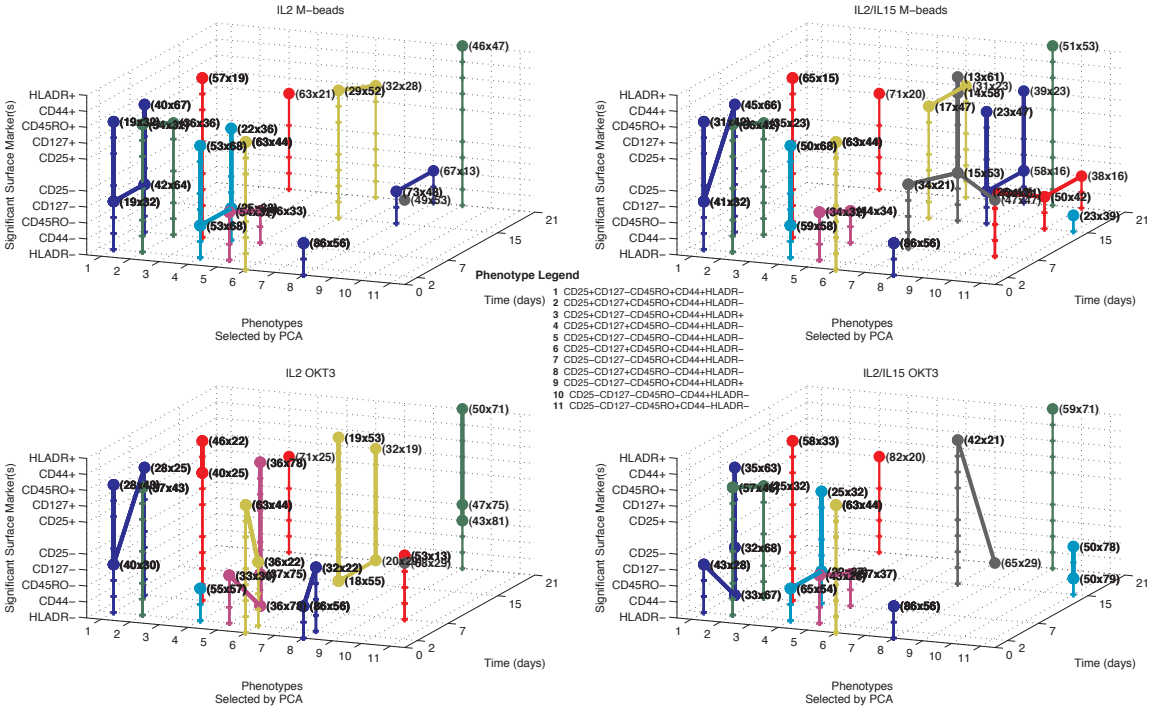


Figure 3-2: Subway plots for the first FACS experiment, performed on CD4 T cells

Group B

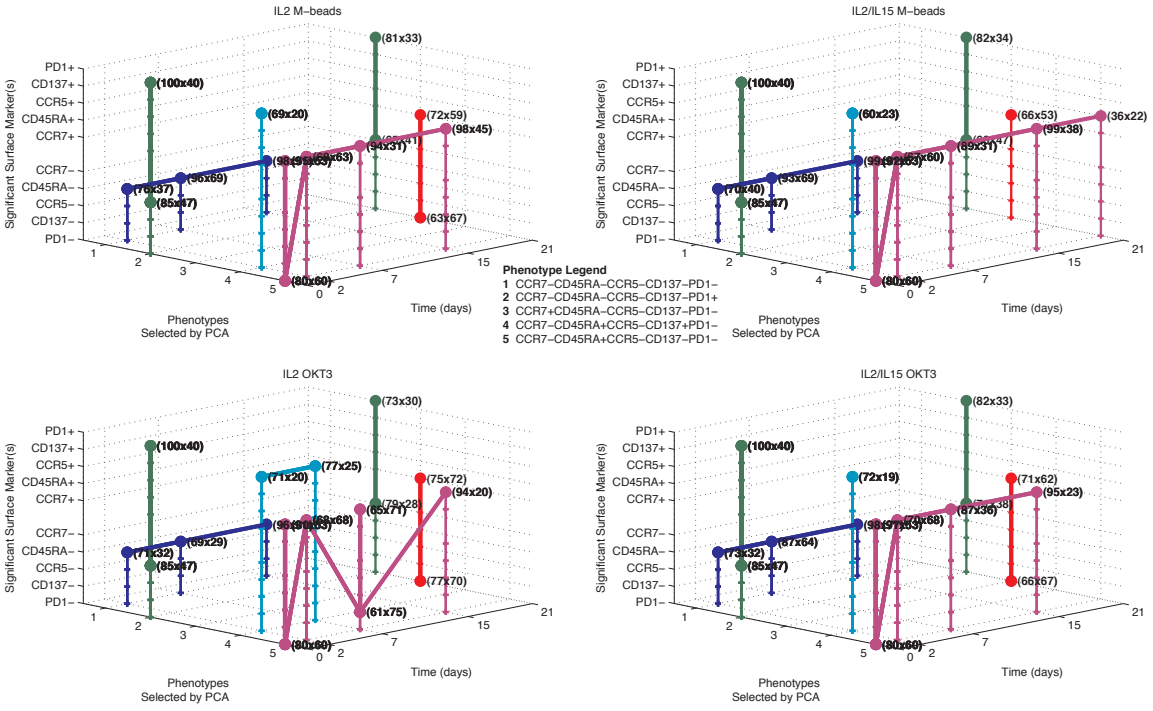


Figure 3-3: Subway plots for the second FACS experiment, performed on CD4 T cells

Group C

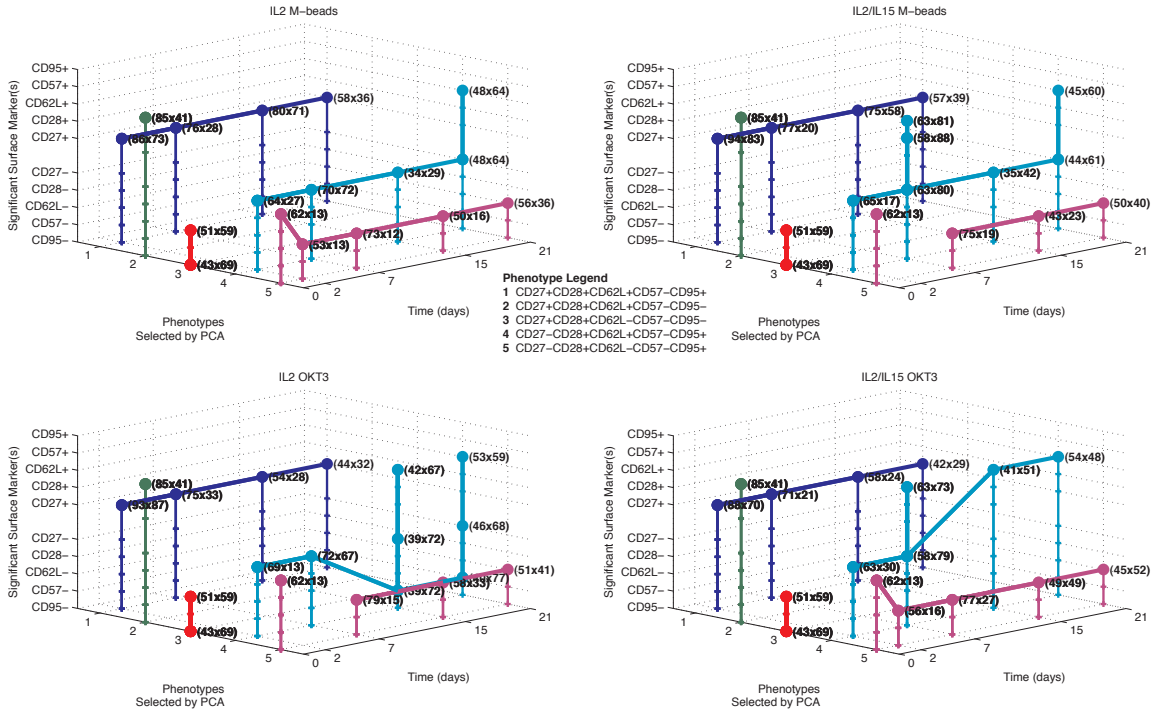


Figure 3-4: Subway plots for the third FACS experiment, performed on CD4 T cells

Group A

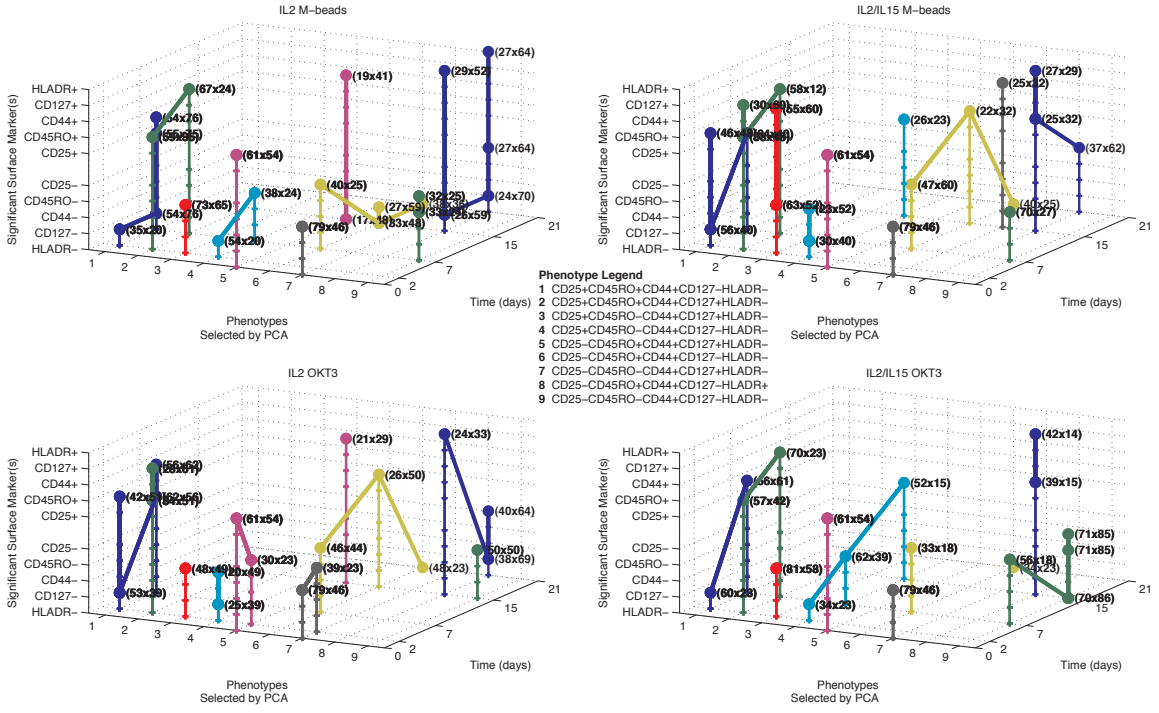


Figure 3-5: Subway plots for the first FACS experiment, performed on CD8 T cells

Group B

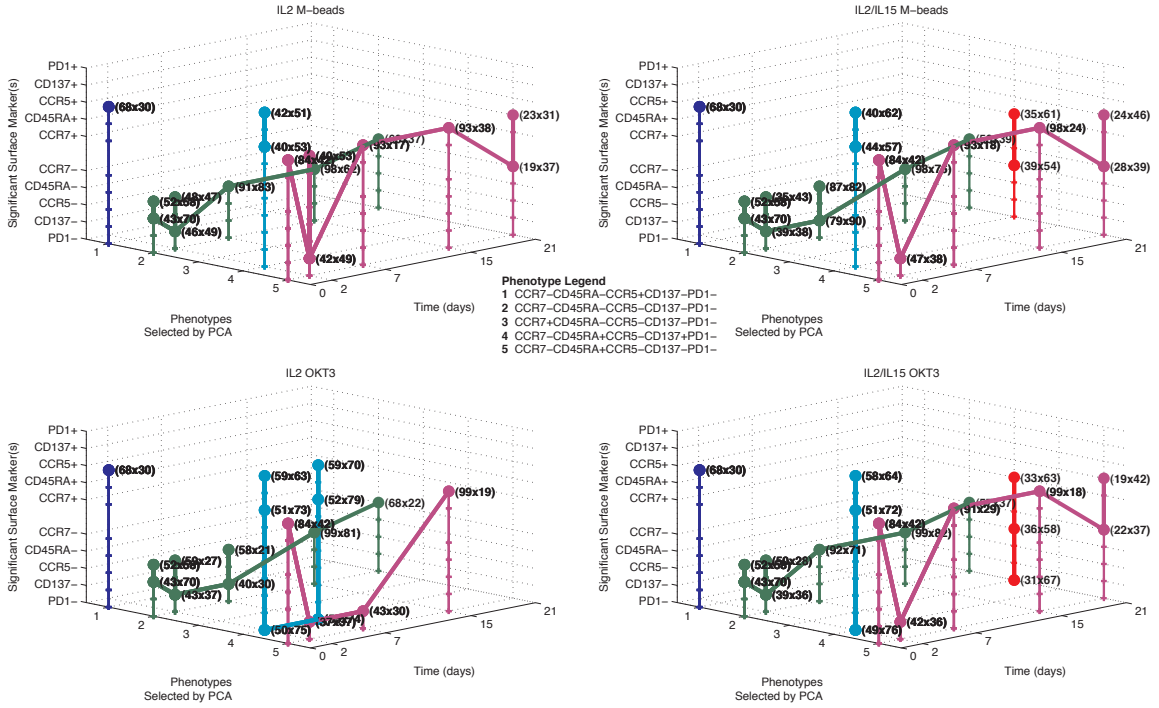


Figure 3-6: Subway plots for the second FACS experiment, performed on CD8 T cells

Group C

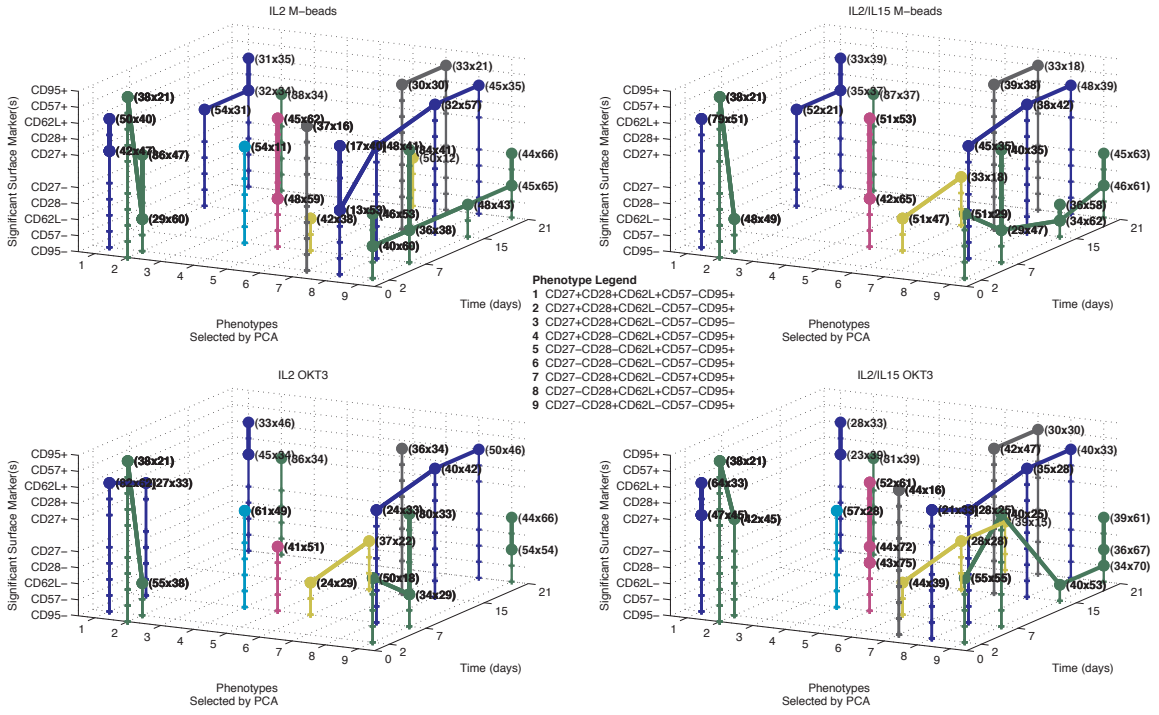
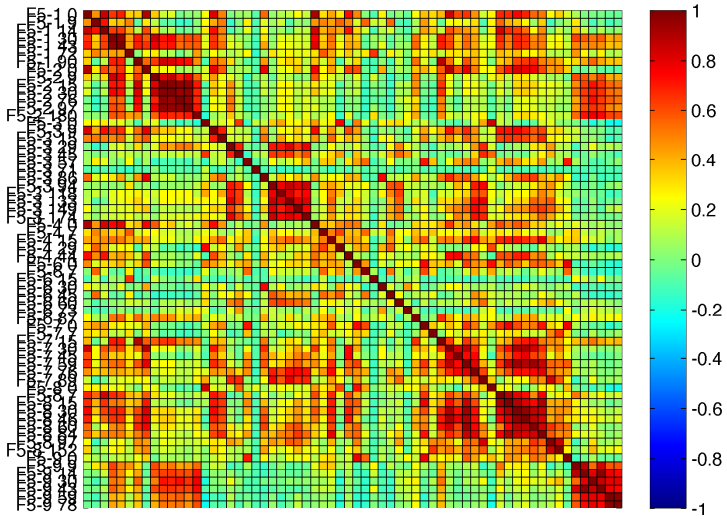


Figure 3-7: Subway plots for the third FACS experiment, performed on CD8 T cells

CD4+ cells



CD8+ cells

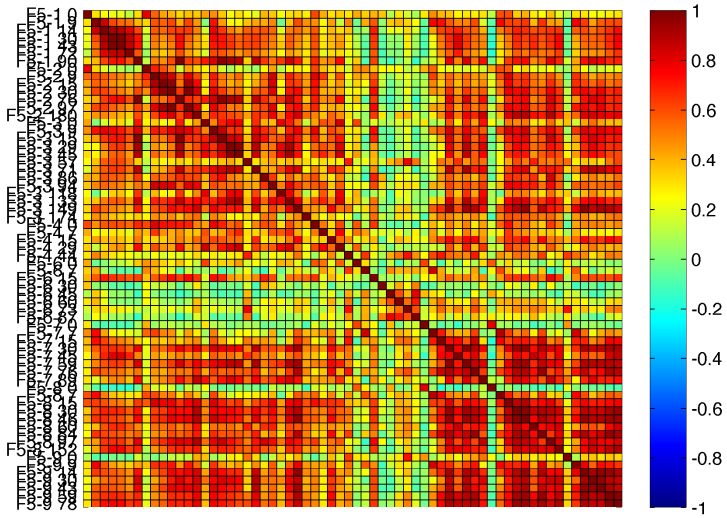


Figure 3-8: Measurement condition correlations highlight the large cross-patient variability of the CD4+ cells and the inconsistency of data from patient F5-6.

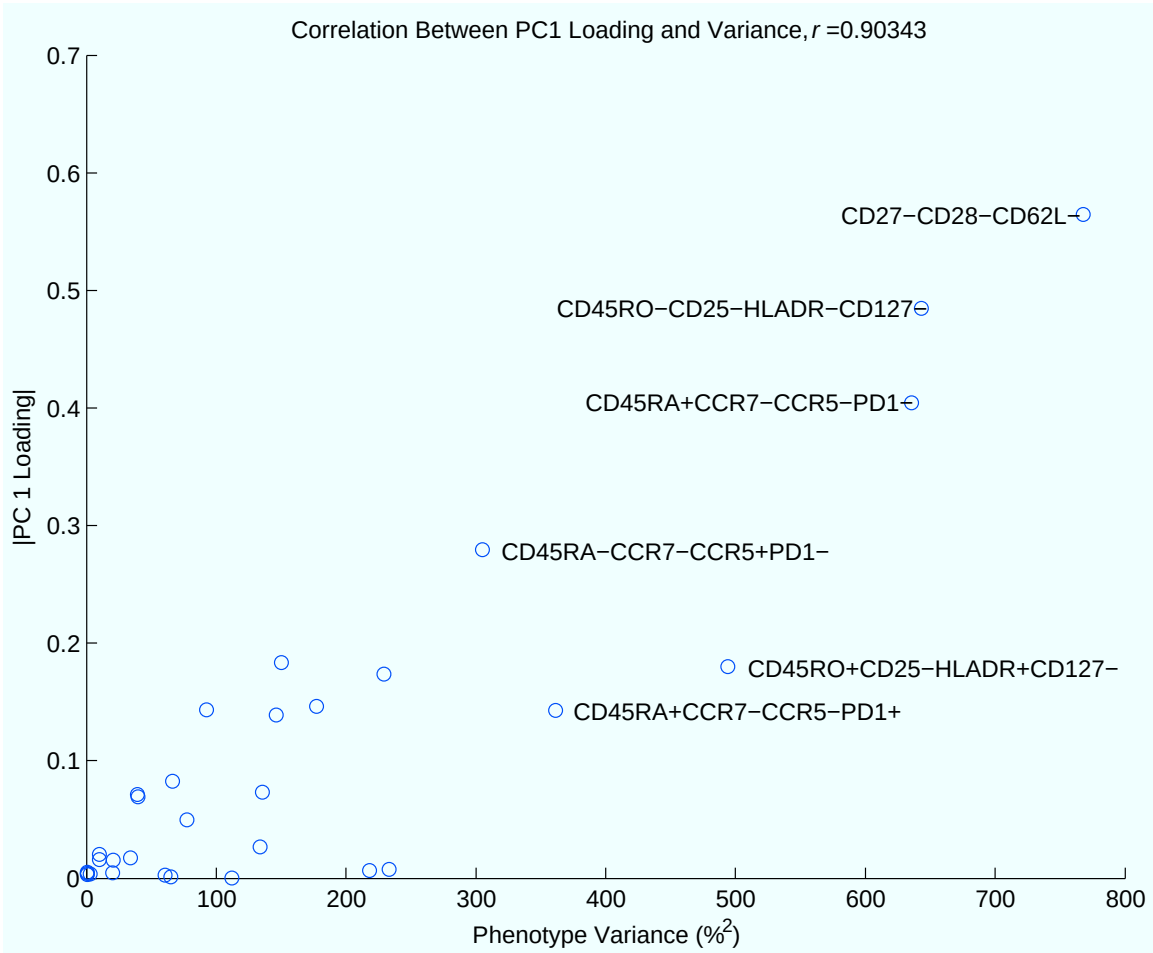


Figure 3-9: Comparing CD8+ phenotypes via their variance and the magnitude of their loading in the first principal component of a covariance PCA leads to a clear separation of a small subset of highly variable phenotypes.

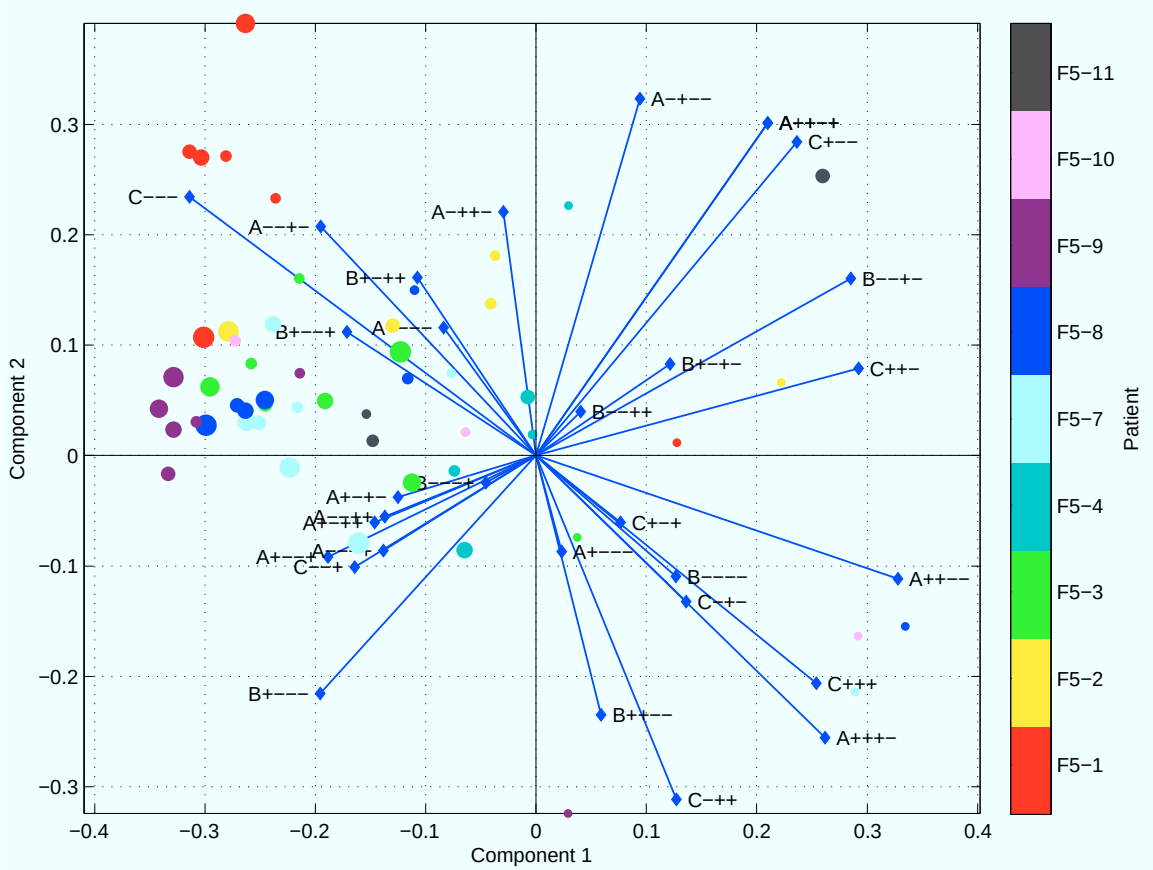


Figure 3-10: A biplot of the data projected onto the first two principal components along with the loadings of each variable for those components identifies the separation of early-time point and late-time point data and the clustering of data by patient. Circle colors represent patient IDs, while circle sizes qualitatively represent the sample day. Phenotypes are identified with a two-part abbreviation. The letter indicates the FACS experiment (and thereby the choice of surface markers), and the pluses and minuses indicate the expression levels of the surface markers that define a particular phenotype. For example, CD27− CD28− CD62L− is identified as C− − −.

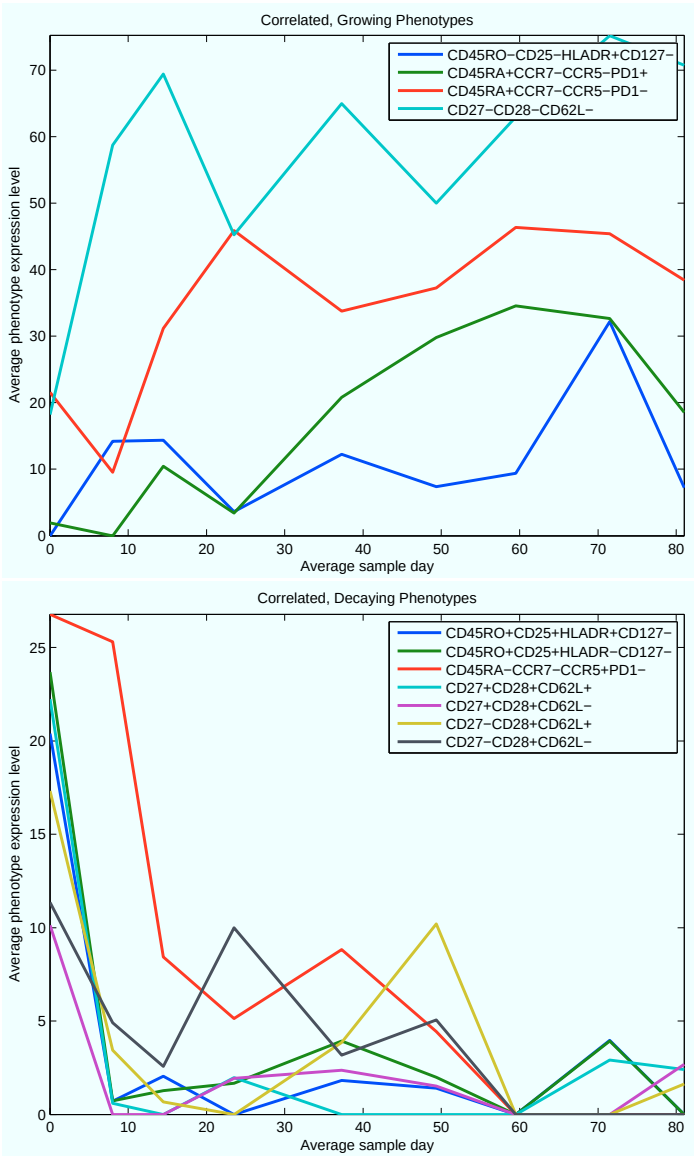


Figure 3-11: A time-based PLS model identifies and separates rapidly growing phenotypes into growing and decaying.

Chapter 4

Differences between blood and spleen cytokine expression levels revealed by a latent-variable regression model

4.1 Summary

The majority of essential immune processes, including antigen presentation, selection, and rest, occur in the tissues. But measurements of the response in the tissues are often prohibitive. Blood measurements, particularly cytokine levels, often serve as a proxy measurement for tissue immune response because they are easy to perform and descriptive. Despite their frequent use as indicators, is unclear how cytokines in the blood relate to the crucial immune activity that occurs in central immune organs. We decided to establish how closely serum and tissue cytokines correlated, using highly multiplexed xMap data from collaborators at Stanford University. The xMap data described levels of 26 cytokines in the serum and the spleen of mice at rest and infected by two strains of *L. monocytogenes*. To be able to understand the highly multidimensional data, we calculated correlations in cytokine levels between serum and spleen cells. The correlations were

surprisingly diverse, so we modeled them with a multidimensional latent-variable regression (O2-PLS). This thesis chapter describes the data collected, outlines in detail the model used to analyze the data, and highlights the following system-level biological insights obtained from the data:

1. the correlations between serum and spleen cytokines are not simply one-to-one

2. using only the correlations between serum and spleen, O2-PLS latent-variable regression models organize the infected cytokine distributions into 4 h. and 12 h. time groups

3. the key cytokines in each time group are distinct and differ between serum and spleen

4. the models also find a common cytokine profile associated with different strains of *L. monocytogenes* infection

By analyzing the correlations between serum and spleen cytokine levels, we constructed a model that extracted biologically relevant information. The highly distinct cytokine profiles expressed by the serum and the spleen suggest that a deeper understanding of the relationship between tissue and serum cytokine levels is necessary for serum measurements to serve as an accurate proxy for the immune tissue response.

4.2 Introduction

One of the main functions of the immune system is to fight pathogens in barrier tissues (e.g., skin, lungs, and gut) and in blood. However, the majority of immune cells are found in central organs such as spleen, lymph nodes and follicles, where

they present antigen, become activated, proliferate and finally rest in wait of another activation signal. In clinical settings, it is easiest to perform immune measurements in blood because it is readily accessible with minimally invasive procedures. Of the many immune molecules that can be measured in blood, cytokines are especially informative direct indicators of immune function. Although blood cytokine measurements are often used as a proxy for tissue cytokine levels, no one has established in depth how the cytokine levels actually relate to one another. Athanassakis and coworkers have suggested that serum cytokine levels correlate directly with spleen cytokine levels [131], but this correlation did not extend to all measured cytokines and was not statistically supported. Nakane and coworkers measured cytokine levels in the serum, spleen, and livers of mice with listeriosis and found substantial differences in the behavior of some cytokines, but their measurements focused only on only three cytokines, and their time-resolution was on the level of days [132]. Because the correlation between spleen and tissue cytokine levels underpins immunologists' understanding and clinical measurement of many immune processes, it must be studied in greater detail.

Through the emergence of antibody-coated bead (xMap) techniques [32], immunologists can now measure many more cytokines simultaneously. xMap beads are internally-dyed polystyrene spheres. Their internal dye mixture allows individual beads to be uniquely identified, and their surface chemistry allows specific binding to a desired analyte. By combining these two independent parameters, xMap instruments detect up to 500 cytokines simultaneously in a single sample with flow cytometry. xMap techniques thus offer the prospect of a deeper understanding of the relationship between spleen and serum immune responses through direct comparison of multiple cytokines from both regions.

Goldberger and coworkers [133] recently collected a novel xMap dataset uniquely capable of addressing the question of spleen-serum correlation (Figure 4-1). They

measured 26 cytokines in the blood and spleen of mice subject to two strains of *Listeria monocytogenes* infection and an uninfected control. Analyzing the correlations in the dataset reveals that the relationship of cytokine levels in the serum and spleen is complicated (Figure 4-1). In comparing serum and spleen cytokine measurements for the same cytokine we might expect one of three outcomes:

1. spleen and serum levels could be positively correlated for the same time point, suggesting both are destinations from a common source.

2. spleen and serum levels could be negatively correlated, suggesting a source where the cytokine travels from one to the other

3. time lapse variations of 1. or 2., suggesting that these events happen in a time delay.

Although some of the above-mentioned correlations do exist in the data, the most common relationship displayed is correlation between *different* cytokines in the serum and spleen. This complexity immediately invalidates the naïve picture of one-to-one correlation between serum and spleen.

The biological picture painted by xMap data in Figure 4-1 presents two distinct analysis challenges: understanding the complex relationship between the cytokine sets, and dealing with the noisy, high-dimensional data. These challenges require that we use multivariate statistical techniques to discern the most informative cytokines and create robust models that make sense of the data. Latent variable models are a simple and powerful type of statistical technique used successfully for many decades in other fields to explore datasets with at least as many dimensions as the cytokine data. [38, 41, 43, 44, 134]. They have recently been applied to problems in biology as well [10, 11, 34, 35, 37, 135]. The fundamental principle common to all such models is to reduce the dimension of a set of data by defining a few new (latent) variables that are weighted sums of the original (manifest) vari-

ables. Calculating latent variables that effectively describe the multivariate data creates a model that is at once more robust and general than naïve multivariate linear regression. Furthermore, the latent variables can serve as useful indicators of important phenomena in the data. Examining the data in the sub-space defined by the latent variables can reveal patterns in the observations, such as a revealing separation or clustering of the data points. Examining the weights that define the latent variables reveals closely associated, correlated groupings of the manifest variables. Often, latent variable models will also relate two blocks of data with a regression. Reducing the data to latent variables makes the regression both robust and general. Latent variable regression thus holds promise for both the noise and complexity challenges posed by the data.

In order to elucidate the relationship between spleen and serum cytokine levels, we constructed a latent-variable model of the data collected by Goldberger and coworkers [133]. Our model reduced the multidimensional data onto a subspace that exposed biologically relevant phenomena. The subspace identified that, in all cases but the uninfected spleen, time was the most important discriminator of cytokine response. Each time point had its own signature cytokine profile, and the cytokines differed starkly between serum and spleen. The latent variables that defined the projection identified the clusters of cytokines associated with each time point. The cytokines grouped by the latent variables were all biologically relevant, even if they did not always belong to a single, easily-identifiable grouping. Many of the cytokines were consistent across different infection strains, identifying a common cytokine profile associated with *L. monocytogenes* infection. The common profile and time separation were not observed in the spleen of the control mice, supporting a role for infection in creating these distinct cytokine profiles.

Our results first establish the complexity of the relationship between serum and spleen cytokine levels and then unravel the complexity by describing the

global features of that relationship. In doing so, they also establish latent variable regression as a useful method for analyzing high-dimensional immune-phenotyping data. The striking differences between serum and spleen cytokine measurements identified by our models establish that serum cytokine levels are poor direct indicators of immune processes in the tissues. Instead, the complex relationship between these two sets of cytokines must be explored with detailed experiments, and predictive models must be established to find a more accurate interpretation of serum cytokine levels.

4.3 Statistical Model

4.3.1 O2-PLS effectively models the cytokine data

As latent variable models are an entire class of analysis technique, we chose a particular member of that class: O2-PLS, a regression technique particularly well-suited for the analysis of the spleen-serum cytokine correlation data [45, 130]. O2-PLS builds a regression that maps two blocks of data to one another—here, the serum and spleen cytokine levels, respectively. The unique feature of O2-PLS is its fundamental separation of variation that is *predictive* between the two blocks and variation that is *orthogonal* to the correlations between them. Predictive variation constitutes the core of the model: it helps explain each block in terms of the other. Orthogonal variation is the compliment of predictive variation: it encompasses any variation that is unique to the individual blocks and is not related to the mapping between them. For modeling biological systems, the presence of the orthogonal group is essential. It allows the model to account for the systematic noise often present in biological data [135]. Explicitly accounting for the systematic noise leads to more interpretable models and allows a deeper understanding of the sources of noise in the data, such as impurities in the compounds or base-

line fluctuations inherent to the measurement [130]. The mathematics of O2-PLS are explained in detail in Section 1.4.

Given training data in two blocks X and Y of equal importance, O2-PLS constructs a model that contains five parts:

1. a low-dimensional latent-variable *predictive* model of the parts of X that correlate with Y

2. a low-dimensional latent-variable *predictive* model of the parts of Y that correlate with X

3. a low-dimensional latent-variable *orthogonal* model of the systematic noise in X, which is completely uncorrelated with Y

4. a low-dimensional latent-variable *orthogonal* model of the systematic noise in Y, which is completely uncorrelated with X

5. a linear regression that removes the uncorrelated parts of X and Y, then symmetrically maps them onto one another.

Each of the low-dimensional models consists of a set of a few latent variables, called components. Each component is a weighted sum of the manifest variables. The weights of this sum are called loadings, and they are constrained so that the sum of their squares is one. The data points, as expressed in the reduced dimensions of the latent variables, are known as scores. The components are defined so that they are completely independent of one another, yet capture as much of the variation and correlation in the data as possible. The first component captures the most variation, and it decreases from there. Although there could be as many components as there are variables, only a small handful are used in practice. In this way, the components provide an effective summary of the most relevant sources of variation in the data. By exploring the regression and studying the scores and

loadings of both the predictive and orthogonal components, O2-PLS gives a thorough description of the correlation and noise in the data.

O2-PLS's explicit separation of predictive and orthogonal correlations provides an effective filter for dealing with noisy biological data, but it also gives the method three distinct advantages over more basic latent variable models such as Principal Component Regression (PCR) [41, 134] and Partial Least Squares Projection onto Latent Structures (PLS, also known as Partial Least Squares Regression) [43, 136].

The first advantage of O2-PLS models is the symmetry between the X and Y data blocks that they impose. Most other latent variable models require distinguishing between a set of independent (X) and dependent (Y) variables. Any regression in these models attempts to establish a one-way relationship *from X to Y*. With the serum and cytokine data, however, one could argue compellingly for either dataset to serve as the independent set: when injecting recombinant cytokines into blood, changes in cytokines can be measured in splenocytes [137, 138], and when introducing cytokines to splenocytes (e.g. through genetic manipulation such as viruses encoding cytokine genes or conditionally expressed cytokines in transgenic mice), one can measure changes in serum cytokines [139]. It is therefore useful not to make any artificial distinction about dependence and instead to treat both sets of data as symmetric blocks that are equally dependent on one another. The symmetric relationship that O2-PLS establishes between X and Y makes it the most logically appropriate choice to model this data.

The second advantage is the improved relevance of the latent variables. By explicitly identifying and filtering out the uncorrelated parts of the data, O2-PLS produces a filtered dataset that decomposes into relevant latent variables that are easier to interpret directly. Without the filtration step, latent variables would incorporate the orthogonal variation into their loadings, making it much more difficult to understand what observations are being distinguished and which manifest

variables are truly relevant. The references that establish O2-PLS [45,130,135] provide a detailed mathematical treatment and examples of this relevance.

The final advantage of O2-PLS follows from the high relevance of the latent variables it produces. In almost all latent variable regression models, particularly PLS, the latent variables obtained are extremely sensitive to the choice of manifest variables used in the data. Even a small percentage of manifest variables that are uncorrelated between X and Y can result in a model whose latent variables are meaningless and impossible to interpret. The need to provide interpretable latent variables imposes an additional problem on such models: variable selection. There are a variety of variable selection techniques, and all require parameterization, either by a cutoff in the correlation value, or a stochastic optimization algorithm, to optimize a measure of model quality across an unfamiliar and rough space [10,47–55]. The sampling algorithms can take a long time to converge, due to the unpredictable nature of the quality statistic, and the exponentially large number of variable combinations precludes any verifiably global optimum solution. In O2-PLS, however, while variable selection can still be performed, it is often unnecessary, as the structured noise contained in the uncorrelated manifest variables is explicitly captured—and analyzed—by the orthogonal part of the model. The orthogonal filtration makes the predictive part of the model more robust to noise and uncorrelated variables, eliminating the need for costly and complicated variable selection.

By removing the necessity of variable selection, O2-PLS serves as a parameter-free technique for understanding the data. The only three parameters in the model are the number of predictive (a), Y-orthogonal (a_{Y_o}), and X-orthogonal (a_{X_o}) components to keep. These parameters, however, are determined explicitly by using cross-validation to calculate the predictive error and finding the optimal numbers of components that minimize this error. In this way, O2-PLS can generate com-

Infection	a	a_{Y_o}	a_{X_o}	${Q^2}_X$	${Q^2}_Y$	$r_{\hat{X}}$	$r_{\hat{Y}}$
none	3	1	2	0.96	0.96	0.81	0.81
L. monocytogenes	5	1	2	0.92	0.96	0.89	0.95
L. m.-L.p.FlaA	3	1	2	0.97	0.96	0.93	0.87

Table 4.1: Cross-validation and goodness of fit statistics show the high quality of the O2-PLS models. The a columns show the number of components in each model. a is the number of predictive components, a_{Y_o} is the number of Y-orthogonal components removed from X, and a_{X_o} is the number of X-orthogonal components removed from Y. $Q^2 \in (-\infty, 1]$ expresses the quality of the model upon cross-validation similarly to how the familiar R^2 expresses the percentage of variance in the original data explained by the model. $r = \sqrt{R^2}$, the Pearson correlation coefficient, is tabulated in the right-most columns for both predictive models. All statistics were calculated from their original definitions for O2-PLS by Trygg and Wold [45].

pletely parameterized models from the data with no arbitrary setting or cutoff parameter.

To model the data from our collaborators [133], we made three separate O2-PLS models: control, wild-type, and mutant—one per infection strain. To determine the dimensionality of the models, we performed the deterministic minimization technique described above on all three models. The resulting validated models all had a low dimension (Table 4.1), highlighting the success of O2-PLS in projecting the 26-dimensional data onto low-dimensional subspaces.

Because of the thorough cross-validation, our O2-PLS models are both parameter-free and high-quality. The models' quality is visible by analyzing statistics, such as R^2 and Q^2, that describe the cross-validation and the prediction (Table 4.1). Only the uninfected model, which has the least potential for meaningful relationships between cytokine levels, has a Pearson correlation coefficient below 0.85, and all models cross-validate well, exhibiting Q^2 values near the maximum of 1, when values above 0.5 are usually considered good [10].

4.3.2 Biplots simultaneously visualize multiple relationships between variables, observations, and latent variables

The key advantage of the O2-PLS models is their ability to turn the unmanageable swath of raw data (Figure 4-1) into a more manageable form. By examining the components of the O2-PLS models, the loadings that define them, and the way the cytokine data are projected onto them, we can turn the original data into meaningful conclusions.

A common visual examination technique for latent variable models is the biplot [39]. Biplots visualize two- or three-dimensional latent-variable representations of a dataset. In O2-PLS, they are used to analyze both predictive and orthogonal models. Each axis of a biplot corresponds to a single latent variable. The biplot displays the relevant information about the latent variables by superimposing two scatterplots, hence its name. The first scatterplot shows the scores—the projection of the data in the subspace defined by the latent variables. Every score corresponds to an observation, and the observations are often identified with relevant information. On top of the score scatterplot sits the loading scatterplot. This plot shows the loadings that define each component. Every point on the loading plot corresponds to a manifest variable, rather than an observation. Loading points are often drawn with lines to the origin for clarity. The scatterplots may seem unrelated, but they share the same dimensions, so the visual proximity they create corresponds to an underlying sense of similarity or relationship.

It is best to ground the abstract description of biplots with a specific example (Figure 4-2). In this contrived dataset, we seek to compare X data of three cytokines (IL-1, IL-2, and IL-3) to Y data of two other cytokines (TNF-α, MIP-1α). The data were designed to exhibit noise, variation in X that is correlated with Y, and variation in X that is uncorrelated with Y. The complete data are plotted in Figure 4-2a, an ideal example of the difficulty of visualizing multidimensional data.

The X data are represented with the three spatial dimensions. The spatial data are composed of two oblate spheroids (a large, tall, skinny cluster and a small, flat, wide cluster) whose major axes point in the same direction, but whose second axes are orthogonal. The Y data are represented with color, with blue representing TNF-α, and pink representing MIP-1α. Purple shades represent a mixture of the two. There is a gradient from pink to purple along the second axes of the smaller spheroid and a gradient from blue to pink along the major axes. Although the data are difficult to visualize, the major axes of the ellipsoids are clearly the largest direction of variance, and they are correlated directly with MIP-1α. The second (flat) axis of the small ellipsoid is coordinated directly to TNF-α, and therefore it makes up the second most useful dimension. Finally, the third axes of both ellipsoids is completely orthogonal to Y. O2-PLS effortlessly detects all three of these features.

O2-PLS allows us to project the data onto a lower-dimensional subspace. Two such possible subspaces are identified in Figure 4-2b. The green plane consists of the first two predictive components. It cuts through the X data along the directions that are directly related to Y. The yellow plane shows that the orthogonal part of O2-PLS finds the uncorrelated direction in X as well. We can analyze both of these projections with biplots, shown in Figure 4-2c – d. The scores of the predictive biplot in Figure 4-2c show the clustering pattern of the two ellipsoids much more clearly than the three-dimensional plot, a "top-down" view. They also outline both directions of variation in Y, since the predictive components correlate strongly with Y. The loadings on this biplot show the orientation of the top-down projection in terms of the original X variables IL-1 – IL-3. Because the scores cluster heavily toward IL-3, we can see that, of the manifest variables, it is the one with the greatest variation. Because IL-3 is plotted exactly along the X axis, the further left a point is along the X axis, the higher its IL-3 value will be. The orthogonal biplot in Figure 4-2d has a similar role. Instead of the top-down perspective, however,

the biplot shows a side view. In the side view, the pattern in Y from MIP-1α (predictive component 1) can still be seen, but the second part of the pattern is gone, because the ordinate is now plotting systematic noise: points along the y axis of the figure vary with no particular order. By orienting this plot by the loadings, we can see that this perspective shows that the axis connecting the clusters at a 45 degree angle in the IL-1 – IL-3 space. It is also consistent with the small flat sphere's higher IL-3 levels, even though the flat sphere is no longer immediately on top of the IL-3 loading. The noise perspective shows that this part of the data is enriched in IL-1 and IL-2 as well. In short, while viewing this multidimensional dataset in full tells us little about its essential features, finding its principal dimensions with O2-PLS reveals useful information about the shape and nature of the data.

4.4 Results

Our model analyzed xMAP data obtained by our experimental collaborators. Goldberger and coworkers [133] performed Luminex xMAP measurements of cytokines in mice following infection with *Listeria monocytogenes*. Mice were divided into three groups: one uninfected control, one infected with wild-type *L. monocytogenes*, and one infected with *L. monocytogenes*-L.p.FlaA—a *L. monocytogenes* strain designed to activate the Nlrc4 inflammasome [140]. Spleen and sera were collected at four and twelve hours post-infection. Samples were taken from four mice per group, and two technical repeats were done in the Luminex experiment.

4.4.1 One-to-one correlation between serum and spleen fails to explain the xMAP measurements

Visual inspection of the spleen cytokine levels plotted against the serum cytokine levels largely invalidates the unbiased assumption of one-to-one correlation be-

tween the spleen and the serum for every cytokine (Figure 4-1). The visual conclusion is supported quantitatively. If spleen and serum cytokines had a one-to-one relationship, we would expect to see self-correlations (diagonal entries in the plot matrices in Figure 4-1) close to unity. In reality, the average self-correlations (diagonal entries in the plot matrices) [141] are 0.15, −0.035, and −0.13, for the no-infection wild-type, and *L. m.*-L.p.FlaA infections, respectively (Figure 4-1). Such low correlations are striking. While they do imply a small tendency toward positive correlation during the normal resting state, this correlation is too weak to support. The slight anti-correlations contradict the one-to-one correlation assumption even more strongly.

The scatterplots in Figure 4-1 also refute the assumption of independent cytokine levels. If serum and spleen cytokines were independent of one another, we would expect the off-diagonal correlations in Figure 4-1 to be vanishingly small. In reality, across all infection conditions, at least 92% of the off-diagonal elements with finite correlations have values outside of the 95% confidence interval of this assumption: a vast majority of the off-diagonal elements signify some measure of correlation between a cytokine in the serum and *other* cytokines in the spleen. The existence of the off-diagonal correlations is consistent with previous results, since since it is known that many cytokines are regulated by other cytokines: indeed, in a systematic immune response, it would be surprising if off-diagonal correlations were absent [142]. Nevertheless, the pervasiveness of the correlations, even among a general set of cytokines without an underlying biological selection criteria, speaks to the complexity of the relationship between spleen and serum cytokine levels. Clearly, the simplifying assumptions of independent, one-to-one relationships between serum and spleen cytokines are invalid. Any model based on them is insufficient to describe the relationship. We must instead turn to a multivariate model that embraces the entire correlation structure: O2-PLS.

4.4.2 Biplots reveal a clustering of the data into 4 h. and 12 h. time points

To understand both the sources of correlated variation in the data and the sources of structured noise, we made biplots of predictive components against orthogonal ones (Figure 4-3). The first predictive component in all three models has positive loadings for all variables. By definition, it reflects the average net correlation between the X and Y data [38]. Because it does not group or distinguish any variables, it is of no great interest to us; for us, the second predictive component becomes the most important. The biplots in Figure 4-3 therefore span the second predictive component and the first orthogonal component.

A pattern emerges from examining these biplots: in the sera and spleens of infected mice, the second principal component clearly separates the data into time-based groups of 4 h. and 12 h. on opposite sides of the origin. As an internal control, we note that uninfected mice do not display this time separation in the spleen, because without infection, immune tissue cytokine levels are in steady state with only small perturbations (Figure 4-4). This separation is particularly vauable, because the model does not initially know anything about time: the separation into timepoints was derived entirely from the covariation of spleen and serum cytokines.

Examining the loadings of the second predictive component (along the abscissa) explains which cytokines are responsible for the time point separation and shows the associations between cytokines and time points (Figure 4-3). In uninfected mice, spleens show no separation between time points, and sera show muddled differences that probably represent steady state changes in cytokines over time. However, after infection with *L. monocytogenes*, we see a clear separation between time points with all the mice in each group driving the differences to varies degrees.

In the uninfected serum, the separation between the time points is mediated by MIP-1α levels at 4 h. and TNF-α levels at 12 h. (Figure 4-3a). In the serum during wild-type *L. monocytogenes* infection (Figure 4-3b), TNF-α is instead associated strongly with 4 h., and IL-5 is the strongest marker of 12 h. In the *L. monocytogenes*-infected spleen, however, TNF-α and MIP-1α are *both* associated strongly with 12 h., while RANTES, TGF-β, and IL-2 are all strongly associated with 4 h. Finally, for the *L. m.*-L.p.FlaA mutant infection (Figure 4-3c), a related set of cytokines emerge. In the serum, TNF-α is strongly associated with 4 h., while IL-5 and TGF-β are associated with 12 h. In the spleen, just like in the wild-type *L. monocytogenes* infection, the 4 h. data is strongly associated with RANTES, TGF-β, and IL-2, while the 12 h. observations are associated with IFN-γ, while MIP-1α plays a secondary role. These results are summarized in Table 4.2, which demonstrates how the correlated groupings of cytokines that O2-PLS detects describe the infections.

The other striking feature of this time-based grouping of cytokines is how starkly different the corresponding cytokines are between serum and spleen. At a given time point, there are *no common cytokines* between the two datasets. The dissimilarity of the spleen and serum cytokine sets is a strong indicator that blood serum measurements do not provide an accurate picture of the current immune response in the tissues. Working out how the cytokine levels in one relate to the other will require additional modeling and measurements. Determining the factors that influence the relationship is a rich biological question requiring a significant effort.

The time separation painted by the biplots is clear, but the association between loading and score is only an implication: there could potentially be hidden variables in the model that cause a score to map closely to a loading. To determine whether or not the associations between time point and cytokine level implied by the biplots are reflected in the cytokine levels themselves, we grouped the cytokine levels by time point (Figure 4-4). This regrouping would not have been

Infection	Serum		Spleen	
	4 h.	12 h.	4 h.	12 h.
none	MIP-1α	TNF-α	–	–
L. monocytogenes	TNF-α $\downarrow$ IL-12p70	IL-5	IL-2 TGF-β RANTES	TNF-α $\uparrow$ MIP-1α IL-1β IFN-γ
L. m.-L.p.FlaA	TNF-α IFN-γ $\downarrow$ IL-12p70	IL-5 TGF-β $\uparrow$	IL-2 TGF-β $\downarrow$ RANTES	IFN-γ $\uparrow$ MIP-1α KC IP-10

Table 4.2: Cytokines with large-magnitude loadings in the predictive component define a common profile of *L. monocytogenes* infection (highlighted in red). Listing the most relevant cytokines together also shows the distinction between serum and spleen, infected and uninfected cytokine fluctuation and allows the identification of temporal patterns in cytokines that appear first in the serum and then in the spleen, or vice versa. The patterns are identified with arrows, which point down to indicate the "source" and up to indicate the "destination."

obvious without the insight provided by the biplots, but it serves as an effective verification of that insight, as the regrouped data exhibit the same separations as in the biplots. In particular, the regrouped plots in Figure 4-4a explain why there is no time-separation in the uninfected spleen biplot: hardly any of the cytokines change with time, so there is no variation to detect. In some cases, such as the infected spleen (Figure 4-4b – c), the 4 h. cytokines and 12 h. cytokines are not inversely related, as one might assume from their separation onto opposing groups. Instead, the distinction between these groups is one of growth: cytokines associated with 4 h. timepoints have moderate to high levels that do not grow at 12 h., while the 12 h. cytokines grow noticeably from their low 4 h. levels. Nevertheless, there is always a clear visual difference between the behavior of both groups, so the highlight plots successfully show that the intuition developed in the biplots is accurate.

4.4.3 O2-PLS clustering reveals the strongest relationships between cytokines in the serum and spleen

In addition to highlighting the stark differences between serum and spleen cytokine levels, the separation of the data into 4 h. and 12 h. timepoints also identifies some of the most significant relationships between serum and spleen cytokine levels. By using the timepoint as a bridge, we can determine relationships between the spleen and serum cytokines during infection. Sometimes, these relationships are not directly played out in a 1:1 correlation, such as the rather weak relationship between serum TNF-α and spleen IL-2, despite both of them being associated with 4 h. time points (Figure 4-1b). Other times, however, they are quite strong, such as the correlation between serum IL-5 and spleen IFN-γ, IL-1β, and MIP-1α (Figure 4-1b – c). Regardless of the strength of the underlying correlation, the spleen-serum associations created by the biplots have meaning because the association based on timepoint is inherently meaningful.

As in the correlations, none of the relationships is a simple one-to-one autocorrelation. This speaks to to the strong differences between the cytokines and the difficulty of extrapolating to tissue immune responses from serum measurements. However, one starting point for identifying a mapping between the two sets of cytokines is a temporal pattern identified by the model. In this pattern, a cytokine is strongly expressed in the serum at one time point, and in the spleen at the other. Each infection variant exhibits at least one such relationship (Table 4.2). In the wild-type *L. monocytogenes* infection, TNF-α is in the serum at 4 h., and the spleen at 12 h., and in the *L. m.*-L.p.FlaA infection, TGF-β from is in the spleen at 4 h. and in the serum at 12 h., and IFN-γ is in the serum at 4 h. and in the spleen at 12 h.. The temporal may not correspond to biological transport, but they do elucidate one of the many complex mechanisms occurring during infection.

4.4.4 O2-PLS models reveal a common cytokine profile of *L. monocytogenes* infection

Cytokine levels in the uninfected spleen seem almost constant over the timescales measured (Figure 4-4). Yet even without infection, cytokines in the serum fluctuate over these timescales. Thus, the presence of cytokine fluctuation alone cannot serve as an indicator of infection. The picture in sera is more complex, because blood is a conduit of cells, proteins, and bacteria. It is a complex representation of events coming from not only the spleen, but also all other organs. However, stepping back from the specifics of the biplots and analyzing them together reveals important general information about cytokine responses in *L. monocytogenes* infection.

Examination of the biplots and the significant cytokines in them (Table 4.2) shows that rather than the presence of variation in cytokine levels, infection is indicated by a cytokine profile common to both infections. The common profile is defined by TNF-α in the serum at 4 h., IL-5 in the serum at 12 h., IL-2, TGF-β, and RANTES in the spleen at 4 h., and MIP-1α and IFN-γ in the spleen at 12 h. (highlighted in red in Table 4.2). These cytokines account for most of the highly-loaded cytokines in the second predictive component of the infected models. The presence of such a core indicates that *L. monocytogenes* infection, even by a mutant strain, induces a well-defined immune response over short time scales. It also testifies to the power of O2-PLS and biplot techniques for detecting this common structure automatically.

The common core of infection also starkly separates infection from the uninfected state. Even if we ignore the idleness of the uninfected spleen, we see that uninfected cytokines vary in different ways from infected cytokines. In the uninfected serum, 4 h. timepoints are associated with MIP-1α and 12 h. timepoints are associated with TNF-α. During infection, however, TNF-α is associated with 4 h.,

while MIP-1α is associated with 12 h., and in the spleen at that, not in the serum. TNF-α and MIP-1α are shown to play a role in both infected and uninfected mice, but their behavior changes. The difference shows how *L. monocytogenes* infection rewires already active signals in addition to introducing new, independent ones.

4.4.5 O2-PLS models reveal strain-specific infection response

Despite sharing a large common core of cytokine activity, the immune responses to the different *L. monocytogenes* strains do express unique features. The mutant *L. m.*-L.p.FlaA infection engages additional serum cytokines: IFN-γ at 4 h. and TGF-β at 12 h. The wild-type *L. monocytogenes* infection, for its part, engages TNF-α at 12 h. in the spleen: a feature absent in the *L. m.*-L.p.FlaA infection.

4.4.6 The orthogonal component in the biplots identifies inter-mouse variation in serum and spleen

In addition to displaying useful information about correlated variation that groups cytokines and observations into relevant clusters, the biplots also inform us about the "structured noise" [45] in the model—variation across mouse or sample that may not have systematic biological meaning but cannot be controlled for. This is intra-block variation local to either the serum or the spleen and unrelated to the correlations between them. The structured noise is plotted along the ordinate in the biplots. In every biplot (Figure 4-3), this axis separates mice from one another, which is completely consistent with its definition: variation that distinguishes mice from one another is noise, uncorrelated between serum and spleen, as all mice were subject to the same treatment.

In the serum, the orthogonal component separation usually distinguishes a particular mouse or two at 4h and shows the cytokines that set that mouse apart. In

the spleen, the orthogonal component typically separates a mouse at 12 h.. For example, in the *L. m.*-L.p.FlaA infection (Figure 4-3c), the serum orthogonal component draws a contrast between Mouse 3, which has relatively higher levels of TNF-α and IL-12p70, and Mouse 4, which has higher levels of MCP-1. Similarly, the spleen orthogonal component in the wild-type *L. monocytogenes* infected-mice (Figure 4-3b) separates Mouse 3, which has elevated levels of the CXC cytokines KC and IP-10, and Mouse 2, which has elevated TNF-α. The continued appearance of TNF-α in both the predictive and orthogonal components highlights a subtle independence in their interrelationship: although the two classes of components are mathematically constrained to be opposites of one another, cytokines can score highly in both of them, in the same sense that a normal distribution can have both a high mean and a high standard deviation without one affecting the other.

The orthogonal, inter-mouse variation also serves to distinguish the *L. monocytogenes* strain infections from one another. In the 12 h. spleen, these differences involve the interplay between predictive variation and structured noise for the cytokines IL-1β, IP-10, and KC. In the wild-type *L. monocytogenes* infection, IL-1β plays a prominent role in the predictive association between serum and spleen and is a reliable marker of the 12 h. timepoint. IP-10 and KC, on the other hand, have only a small correlated contribution and instead have sizable orthogonal loadings that indicate they are a key source of mouse-to-mouse variation. In the *L. m.*-L.p.FlaA infection, this dichotomy is changed. IL-1β instead makes hardly any predictive contribution, and is instead strongly associated with the orthogonal, structured noise in the spleen. IP-10, on the other hand, is now solidly predictive, and has hardly any orthogonal loading at all. KC-*h*as also become predictive, but still maintains a strong orthogonal contribution, and MIP-1α, while always being a core predictive cytokine, now also has a significant orthogonal contribution as well. It is likely that some of this interplay between predictive and orthogo-

nal variation is due to the relatively small number of mice included in the cohort. If the number of mice were greater, a more consistent picture of structured noise could emerge. Nevertheless, the current O2-PLS models succinctly identify the general core features of *L. monocytogenes* infection but are still capable of capturing differences between variants.

Exploring and analyzing the inter-mouse variation is a sizable advantage for O2-PLS over methodologies such as PLS with variable selection: variables that have both a high predictive contribution and a high orthogonal contribution can be left in the model to provide useful information rather than being discarded simply to reduce noise. The contribution of those variables to the structured noise may also be used to distinguish experimental conditions or to understand what experimental parameters are most important to control.

4.5 Discussion

In this chapter, we have examined the nature of cytokine interactions between spleen and serum using a powerful statistical technique, O2-PLS. We have established that this is a biologically relevant question that can now be addressed much more thoroughly with new xMAP techniques. We have explained the features of O2-PLS that make it an ideal choice for analyzing such data. And, using data from recent experiments on *L. monocytogenes* infection of mice, we have demonstrated the power of O2-PLS to extract essential information from an overwhelming wealth of cytokine levels. Our O2-PLS models are robust. We used them to identify the most relevant cytokines, show how cytokine levels in the serum related to those in the spleen, cluster the observations into meaningful time-based groups, and learn a common temporal and molecular signature of *L. monocytogenes* infection.

The most striking feature about the data and its analysis is the stark differences

between serum and spleen cytokine levels that the model identified. These point to a highly complex relationship between the two systems. To conclude, we suggest some potential biological motivation for the most prevalent cytokines in our model. These hypotheses represent a starting point for understanding the relationship between serum and spleen through experiment.

When combined with intuition and prior knowledge of the pathology of listeriosis, the correlated groups of cytokines can be interpreted as hints about a mechanistic understanding of *L. monocytogenes* infection, subject to the limits of the data. Since the data were only collected at 4 h. and 12 h. timepoints, it seems likely that the cytokine levels represent mostly inflammation and other innate immune responses, rather than a strong adaptive response, as *L. monocytogenes* has barely begun to enter the T cell zones by 12 h. [143, 144]. Many of the cytokines highlighted by the model as being in the core of infection are often associated with innate responses to *L. monocytogenes* [143]. Based on previous research, we can hypothesize a likely underlying mechanistic cause for each core cytokine. While the hypotheses do not lead to a direct mechanistic picture themselves, they are a wellspring of motivation for further experimental study.

One of the most highly expressed and widely-variable cytokines in the infectious core was TNF-α. According to a widespread body of literature, this cytokine is essential for a successful *L. monocytogenes* response [143, 145]. It is strongly associated with macrophage expression and has recently been shown to associate with dendritic cells as well [146]. Given its prominence in the acute inflammation response, it is no surprise to see TNF-α throughout the models, likely indicating macrophage and dendritic cell activity in the serum at 4 h., and in the spleen at 12 h. This activation seems to be tempered somewhat by the presence of TGF-β. TGF-β is known to suppress many of the effects of macrophage activation [147]. It is reasonable to expect its more pronounced prevalence in the *L. m.*-L.p.FlaA

infection, which tends to be weaker. Finally, MIP-1α, another key component of the infectious core, is produced by macrophages, as its name implies, and clearly indicates their presence in the spleen at 12 h..

A second set of cytokines, consistently identified by the model in the serum at 4 h. is IL-12$p70$ and, in the case of *L. m.*-L.p.FlaA, IFN-γ. These cytokines are strongly associated with NK cells. IFN-γ is an inflammatory cytokine that is produced by activated T and NK cells and activates macrophages (among its many functions). It is frequently cited as an essential cytokine for a successful host defense [143, 145], and becomes even more important later on during T cell responses. Its presence in the serum at 4 h. is likely due to its innate immune effects, while its presence in the spleen at 12 h. may already be due to T cell activity [143, 145].

Both TNF-α and IL-12$p70$, which are expressed at 4 h., are involved in lymphocyte proliferation and differentiation. The later is a chemokine that specifically attracts monocytes, memory T-helper cells and eosinophils. Therefore, a combination of control of migration and differentiation is revealed in the spleen at 4 h.. At 12 h., based on the expression of IFN-γ and MIP-1α, the spleen is a site of active inflammatory responses, driven by macrophage-related cytokines. This is in agreement with the role of macrophages and dendritic cells as the host cell for *L. monocytogenes*.

Another uniquely strong signal in the model was IL-5. Its main associations include Th2 CD4$^+$ T cells, but it is also associated with mast cells [148], which have been shown to play an important role in early *L. monocytogenes* infection [149].

IL-6, known to be associated with TNF-α in triggering the acute stage of the response [145], is often seen as a key marker of *L. monocytogenes* infection. In our observation, it was weakly, but consistently observed in the serum at 12 h. of infection. However, since liver samples were not measured in this dataset, and

thus Kupffer cells, a major source of IL-6 [145], were not incorporated, high levels of IL-6 were not observed, and it was not among the strongest signals.

The association of IL-1β, a proxy for inflammasome activation [140] in the model was surprising. Despite Sauer et al.'s observation that wild-type *L. monocytogenes* released low IL-1β levels [140], the cytokine was still identified as a relatively reliable indicator of wild-type infection in the spleen at 12 h. In the *L. m.*-L.p.FlaA infection, designed and observed to elicit strong inflammasome activation [140], IL-1β levels varied highly from mouse to mouse and were thus associated strongly with the orthogonal component, rather than with the predictive model. The models' classifications are not inconsistent with experiments, as they address different properties: IL-1β's grouping depends less on its overall level than it does on its variability across time and across mouse. In this context it is also important to note that IL-1β is very hard to measure in sera and spleen (due to poor reagents) and that the AIM2 inflammasome can be activated by DNA from dead *L. monocytogenes* and confound the observation [150]. Since our experiments did not include a test for IL-18, another cytokine that is secreted as a result of inflammasome activation, we cannot specifically address the magnitude of inflammasome activation in this experiment.

The biological associations implied by the detection of a variety of cytokines are useful indicators of future experimental hypotheses. There are also other directions for extending this work. Most prominent among these is extension along the time axis. Since adaptive immune responses are much stronger at 24 h. and beyond, but *L. m.*-L.p.FlaA is cleared by 48 h., an O2-PLS model of data at such timepoints up to 48 h. would be an even richer, more valuable compass in the forest of analysis. Furthermore, there are other biomarkers which would be fascinating to include in the model—from intercellular proteins such as MyD88, known to be essential in macrophage regulation [143], to Toll-like surface receptors. Gene

expression data for key proteins could also be a useful tool. The ability of O2-PLS to successfully juxtapose and combine multiple kinds of data into a single, all-encompassing model would make it an essential tool in these analyses.

Thus, O2-PLS is a powerful latent variable regression technique that can take large, systematic datasets and extract meaningful information from them, suggest associations, cluster data, motivate mechanistic understanding, and suggest future experiments. Despite its lack of intense computation and its underlying simplicity, the technique holds tremendous promise in biology, and we hope future immunologists will consider its use to bring clarity and order to complex, multidimensional datasets.

4.6 Methods

4.6.1 Data collection

The data used in our analyses [133] was collected by the Stanford Human Immune Monitory Center using Panomics beads and Luminex 100 IS or Luminex 200 machines. Cytokine levels in the spleen and serum were collected using antibody-coated, fluorescent (XMAP) beads [32]. Spleen and serum data were collected on separate plates.

4.6.2 Data normalization

Because of differences in background fluorescence levels between the two plates and among the 26 different cytokines, the data were normalized as follows:

1. Levels of each cytokine were rescaled relative to the MFI at the center of the concentration curve

2. A center value was calculated for each cytokine using the rescaled MFI values

3. A background cutoff value was calculated as 130% of the current background levels. The 130% scale factor was determined visually as the most conservative estimate of MFI levels at which spurious background signal would not be accidentally detected

4. The MFI values from each plate were shifted symmetrically so that the center values for each cytokine were equal between the two plates

5. Any MFI values below the background cutoff were set to zero

6. An outlier-less 100th percentile was calculated for each cytokine, with outliers defined as any MFIs greater than $\frac{3}{2}$ the interquartile range

7. Every cytokine's MFIs were rescaled by the 100^{th} percentile

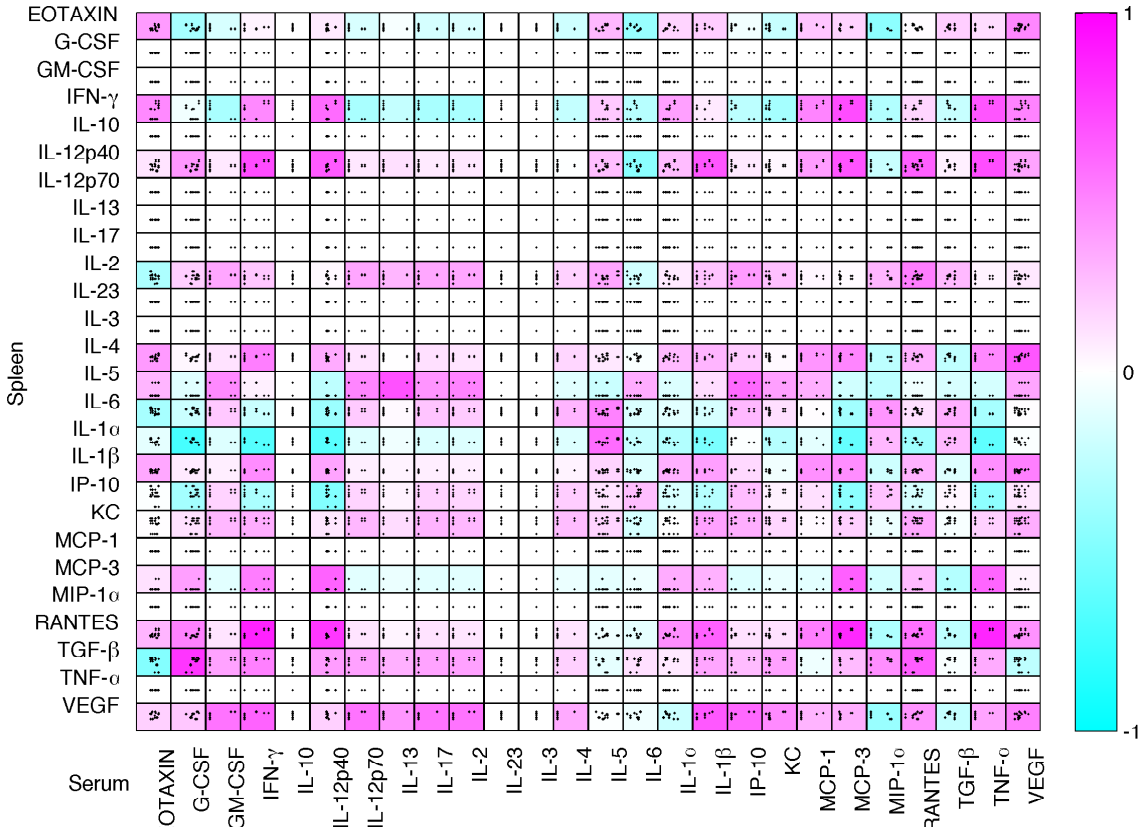
The normalization ensured that MFIs obtained on one plate could be compared to those from another plate, that the data was not biased toward cytokines with better antibodies, and that spurious MFI signals were not treated as data.

In every model, samples with known extenuating experimental issues were removed (both samples per mouse). Finally, for all three models, any variable which fewer than two nonzero data points for the samples in the model was excluded from the model. This simple preprocessing is a sanity check, as compared to the much more complex variable selection preprocessing procedures required for other kinds of regression models [10, 47–55].

4.6.3 Statistical modeling

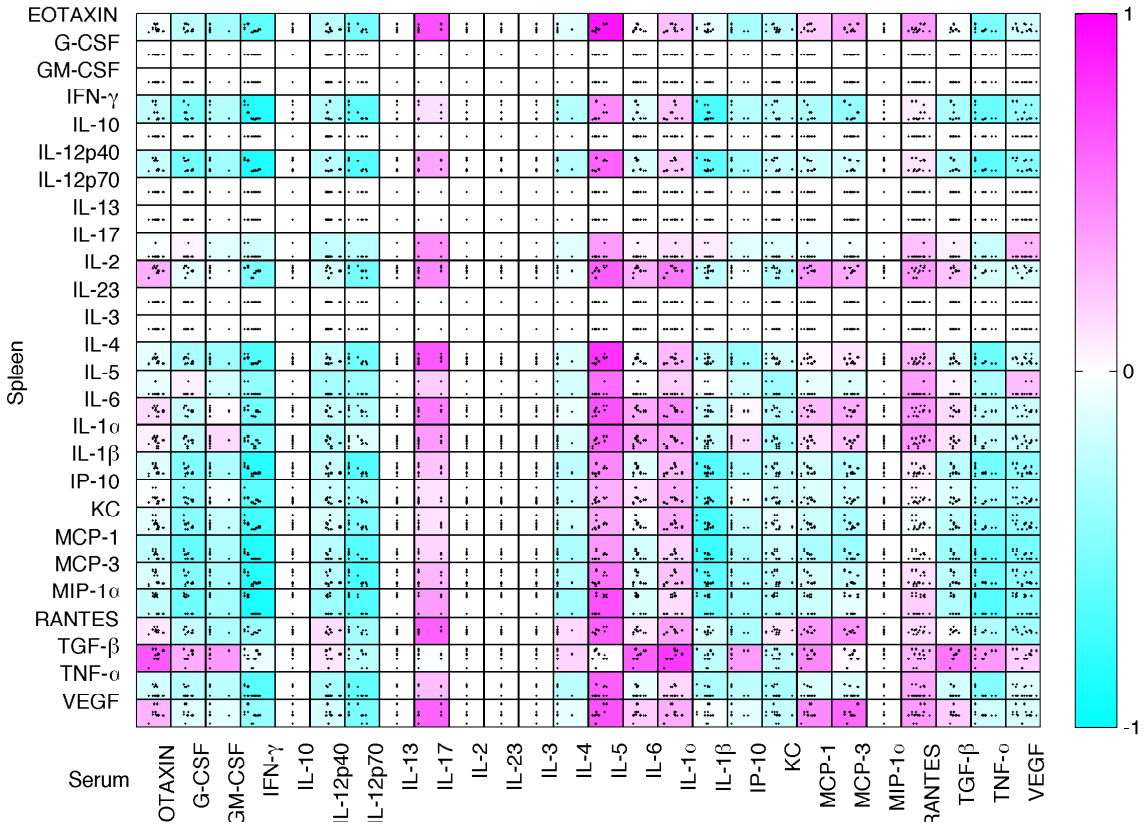
The normalized and filtered data was sample-replicated. Since each mouse contained two independent samples, both for the serum and the spleen, the data was replicated so that each serum sample was paired with both spleen samples, and *vice versa*. O2-PLS models of the replicated data were constructed using software

written from scratch in MATLAB [151], based on algorithms outlined in the papers that defined the method and its extensions [45, 57, 59, 130]. Based on previous approaches to choosing the optimal number of components [45, 152], the optimal number of components was chosen for each model using leave-one-out (LOO) cross-validation and a steepest-descent minimization scheme. First, the LOO mean-squared error (MSE) of cross-validation was calculated for models with all valid combinations of predictive, X-orthogonal and Y-orthogonal dimensions a, $a_{Y_o} \leq a$, and $a_{X_o} \leq a$, producing a three-dimensional grid. The error minimization was then performed in two stages. In the first stage, the optimal a was chosen by calculating the median MSE for each a across all valid combinations of a_{Y_o} and a_{X_o}, and applying a steepest-descent search, starting at $a = 1$. Once the optimal a was fixed, the optimal a_{Y_o} and a_{X_o} were chosen by steepest descent minimization of the MSE starting at $a_{Y_o} = 0$, $a_{X_o} = 0$. The minimizations converged rapidly to local minima. The optimal a, a_{Y_o}, and a_{X_o} found by the minimization scheme agreed with the values chosen by the scree plot method [57, 58].



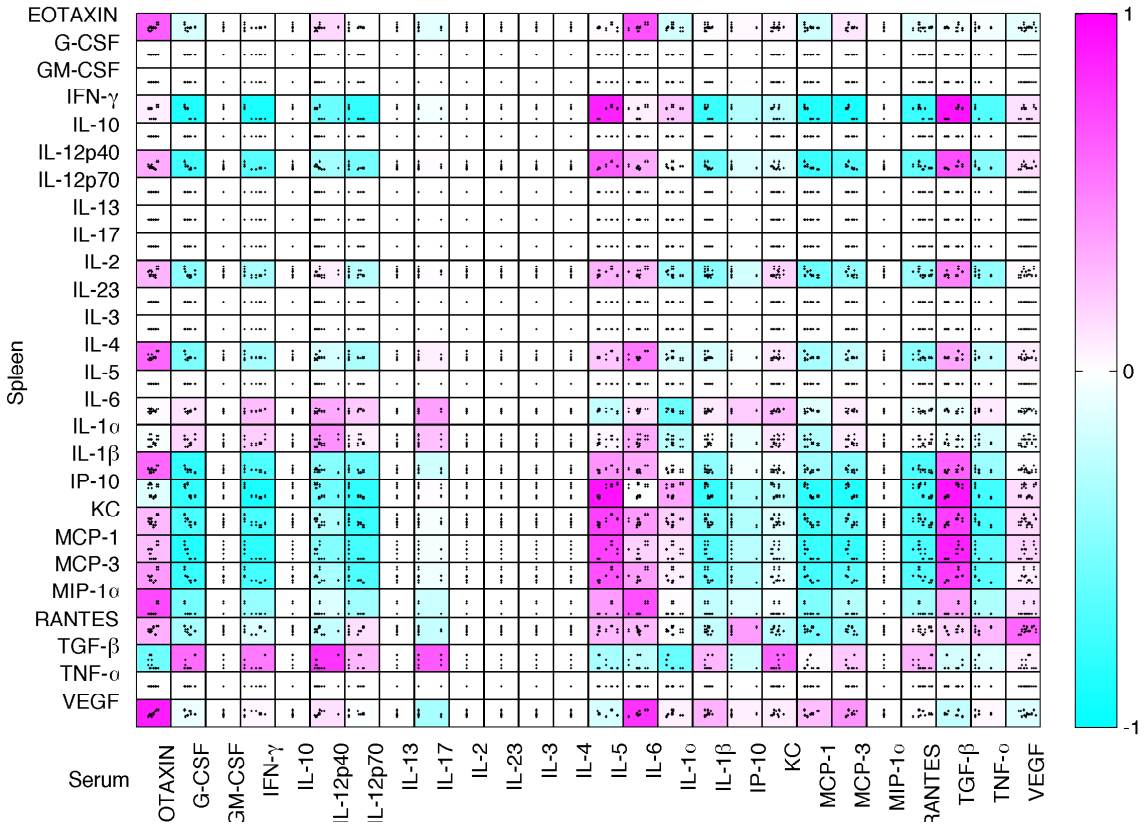
(a) No Infection

Figure 4-1: Plotting the serum cytokine levels against the spleen cytokine levels demonstrates the complexity of the correlations between the two. Each plot represents the levels of one cytokine in the serum plotted against the corresponding cytokine in the spleen, as a function of time. Its background color represents the value of r, the Pearson correlation coefficient for the data. Rarely is a cytokine in the serum strongly correlated to itself in the spleen; sometimes, it is *anti*-correlated. Furthermore, there are substantial off-diagonal correlations between groups of cytokines. A latent variable regression model, such as O2-PLS, tames this complexity.



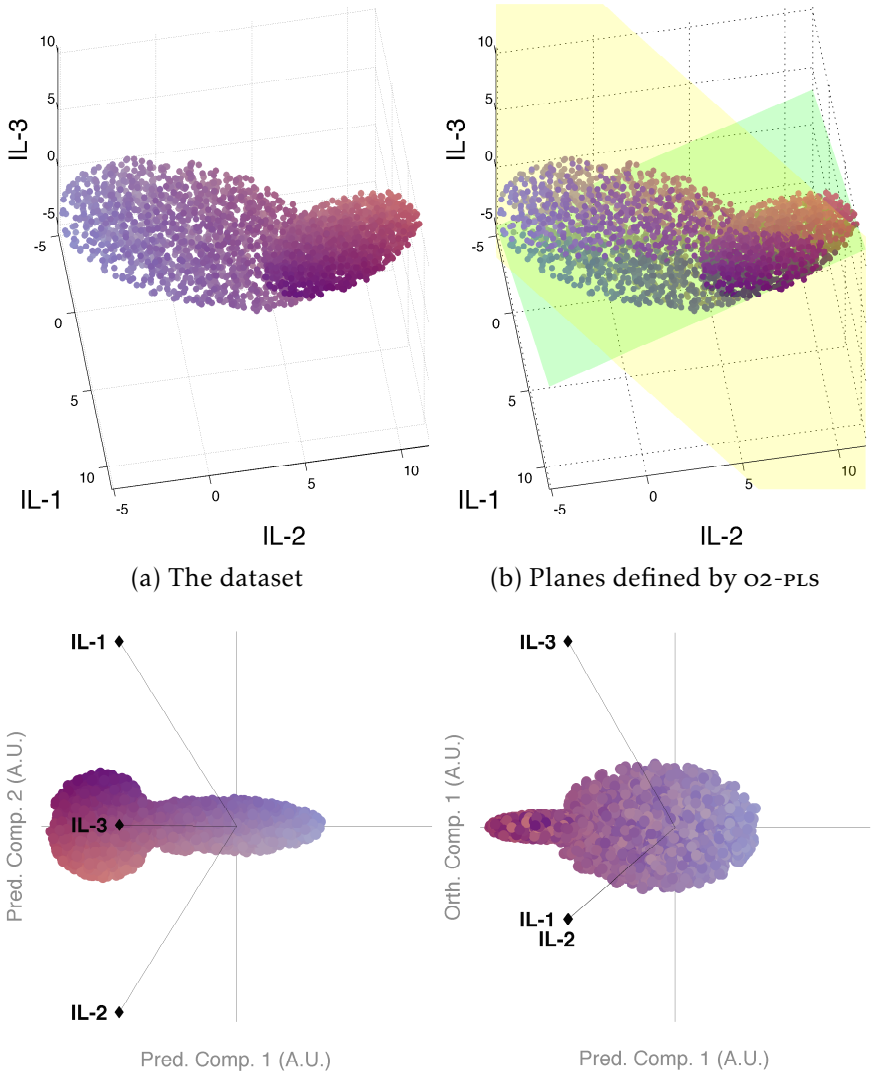
(b) *L. monocytogenes* Infection

Figure 4-1



(c) *L. m.*-L.p.FlaA Infection

Figure 4-1



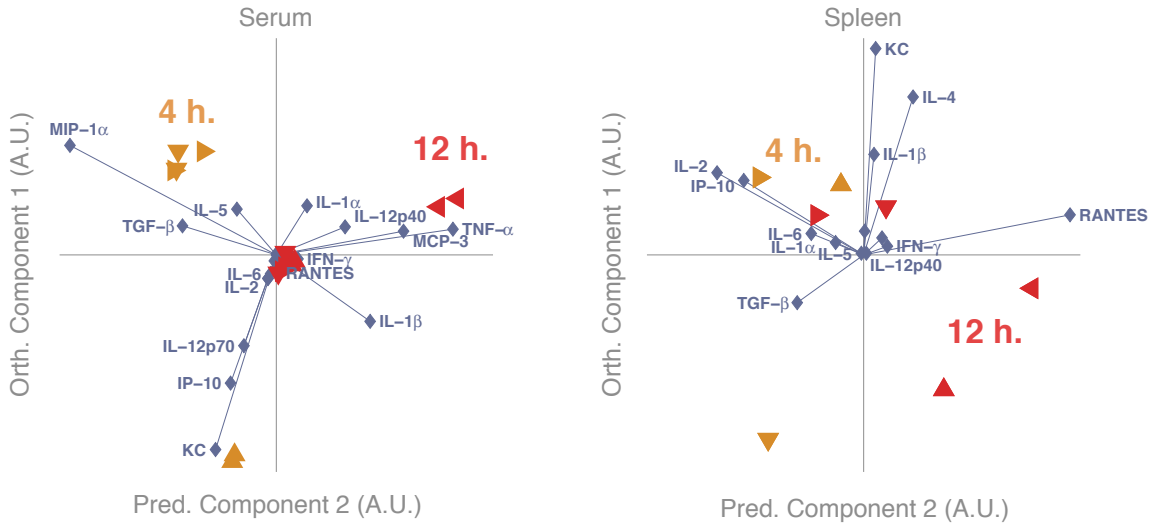
(a) The dataset

(b) Planes defined by O2-PLS

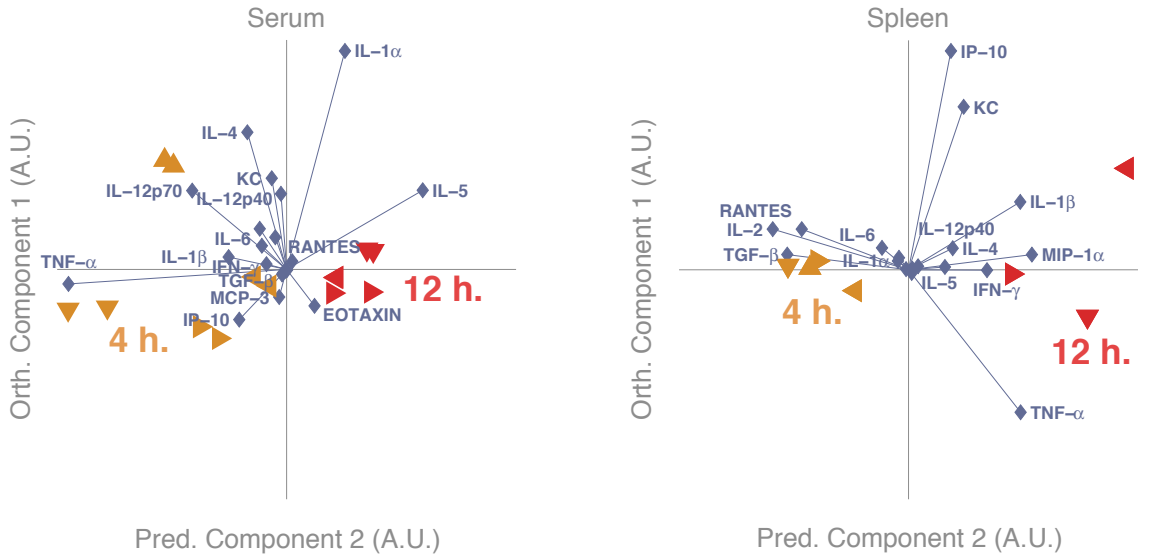
(c) The predictive biplot

(d) The orthogonal biplot

Figure 4-2: An example dataset explains the power of O2-PLS for modeling high-dimensional data and describes biplots. (a) The original dataset is difficult to visualize. The three X-dimensions IL-1, IL-2, and IL-3, are represented spatially, and the two Y-dimensions TNF-α and MIP-1α represented with color. The Y data are plotted as a color mixture, with blue representing TNF-α and pink representing MIP-1α. (b) O2-PLS finds reduced-dimensional subspaces onto which the data can be projected, using both predictive and orthogonal data. The predictive subspace is green, while the orthogonal subspace is yellow. (c) A biplot in the space of the first two predictive components identifies clear trends in the data and explains those trends in terms of the X-variables. Both dimensions of Y are projected cleanly. (d) A biplot in the space of the first predictive and first orthogonal component. The vertical dimension of the pattern in the Y data is no longer visible.

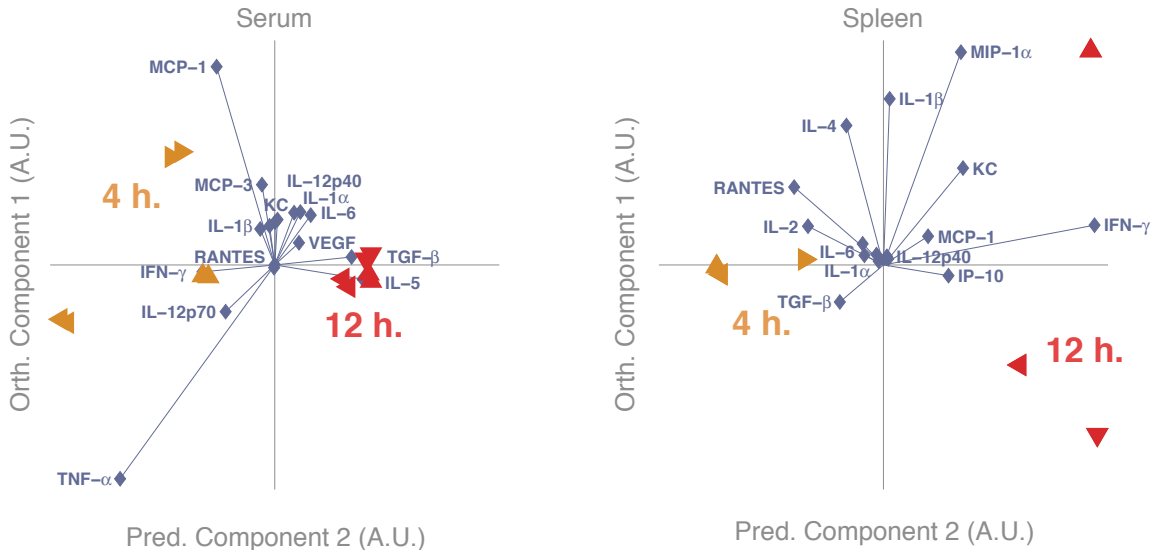


(a) No Infection



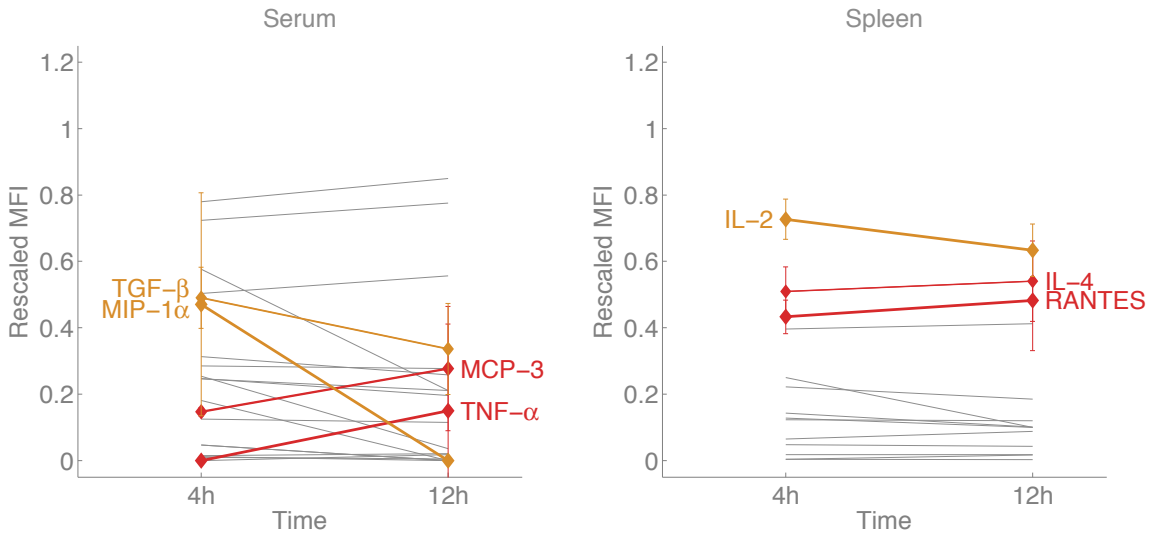
(b) *L. monocytogenes* Infection

Figure 4-3: Biplots of the O2-PLS models highlight the separation of the data into 4 h. and 12 h. groups that differ between serum and spleen, identify the cytokines most relevant to the spleen-serum relation, highlight a common cytokine signature for *L. monocytogenes* infection, and identify the cytokines responsible for differences between infection variants. They simultaneously show the impact of each cytokine on separating the data into time clusters (abscissa), and the impact of each cytokine on the correlated noise (ordinate). The rotation of the triangles represents the mouse ID of a sample: Mouse 1 = △, Mouse 2 = ▽, Mouse 3 = ◁, Mouse 4 = ▷.

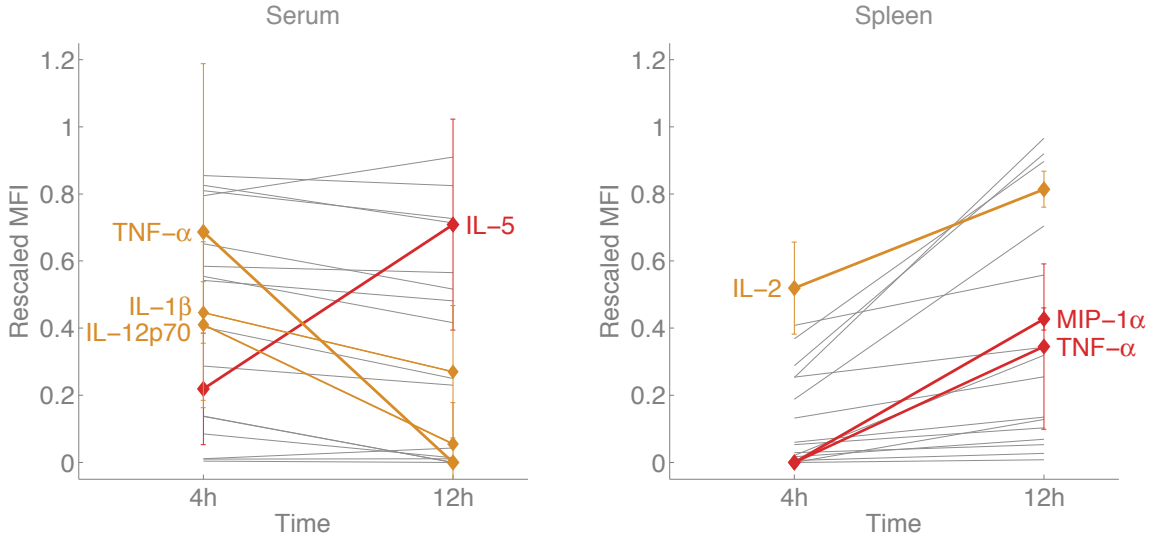


(c) *L. m.*-L.p.FlaA Infection

Figure 4-3

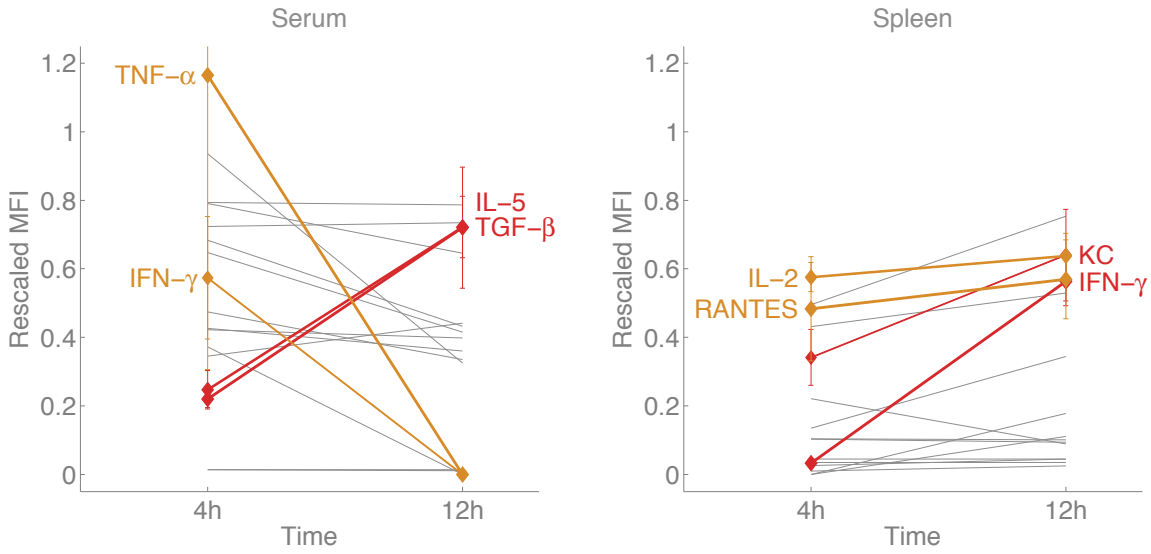


(a) No Infection



(b) *L. monocytogenes* Infection

Figure 4-4: Grouping the MFI data by time-point shows the differences in cytokine level that define the time separation. The plots show MFI level as a function of time. Cytokines associated in the biplot with 4 h. time points (yellow) tend to either have higher levels at 4 h. than at 12 h. or not grow as rapidly by 12 h. as cytokines associated with the 12 h. time points in the biplot. The plots also explain why no time separation was found for the uninfected spleen: its cytokine levels remain virtually unchanged across time. Only cytokines with a loading magnitude greater than the 80^{th} percentile are highlighted, while the rest are gray. The thickness of a line denotes the relative magnitude of that cytokine's loading in predictive component 2.



(c) *L. m.*-L.p.FlaA Infection

Figure 4-4

Appendix A

Supplementary information for Chapter 2

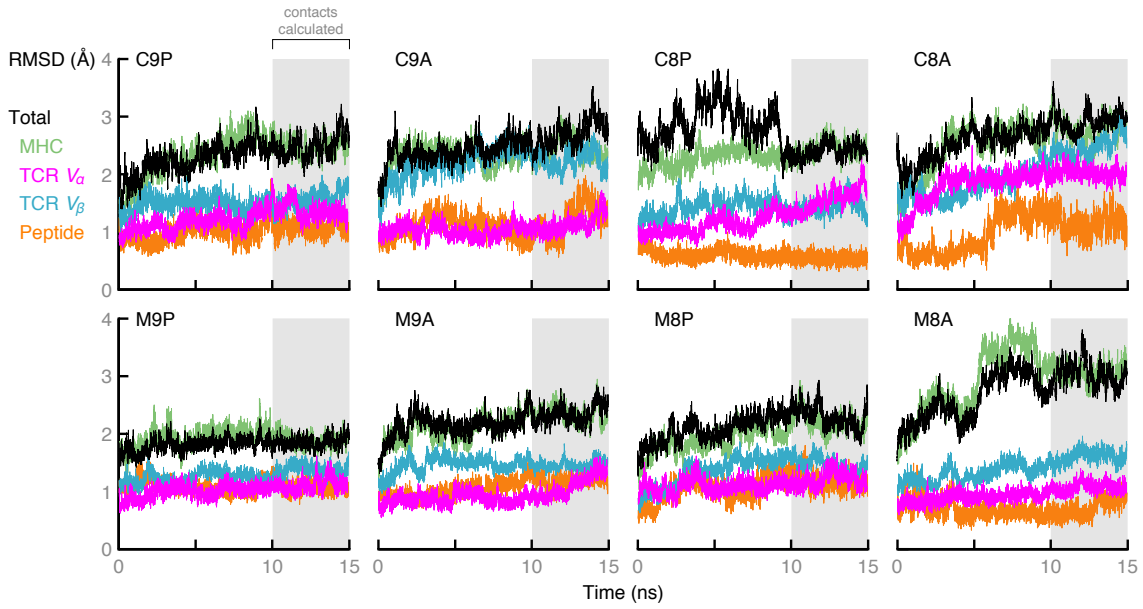


Figure A-1: Backbone RMSD calculations show that the molecular dynamics simulations are stable. The plots show the RMSD of backbone C_α, C, N, and O atoms from the original crystal structure coordinates, aligned to compensate for rotation and translation. In addition to the total RMSD, the RMSD of each constituent protein segment is also shown. The region of trajectories used in the contact footprint analysis (between 10 ns and 15 ns) is highlighted in gray. In this region, the total RMSD values are relatively plateaued. RMSD values after equilibration are generally high, especially for the mutant peptides, but this mostly reflects the fact that the original crystal structure is a poor reference for the equilibrated mutant peptide structures, as it does not take into account any conformational changes necessary to accommodate the new peptides. Furthermore, most of the large RMSD deviations are due to the MHC, particularly the α_1 and α_2 helices (data not shown), which are responsible for maintaining the peptide groove.

Table A.1: Calculated binding free energies are consistent with experiments

TCR	pMHC	Simulation ΔG (kcal/mol)[1]	Experimental ΔG (kcal/mol)[2]
2C	QL9-L^d	-3.2 ± 11	-7.6 ± 0.07
m6	QL9-L^d	-24.7 ± 14	-9.0 ± 0.2

[1]Binding free energies (ΔG) calculated from the molecular dynamics trajectory are qualitatively consistent with experimentally measured binding free energies. Simulation binding free energies were calculated using the MM-GBSA method, [153, 154], averaged over the last 5ns of the trajectory. The MM-GBSA protocol was adapted from the TCR/pMHC MM-GBSA simulations of Zoete et al. [29], with the exception of the entropy term. This term is likely to affect the actual values of ΔG obtained but is not expected to qualitatively change the order of the binding affinities. The comparison shows that our simulations reproduce the experimental finding that m6 binds more strongly to QL9-L^d than 2C. Errors are s.d.

[2] Experimental free energies were obtained by Jones et al. [104] Errors are s.d.

Table A.2: The changes in TCR/MHC contacts upon peptide mutation from C8P to C8A are numerically significant

System A	$\bar{r}_c(A;\alpha_1)^1$	$\bar{r}_c(A;\alpha_2)^1$
C9P	71.4 ± 0.8	156.1 ± 0.9
C9A	72.4 ± 0.8	156.0 ± 1.0
C8P	69.5 ± 0.8	156.7 ± 0.7
C8A	73.5 ± 1.0	153.5 ± 0.8
M9P	71.1 ± 0.8	156.7 ± 0.6
C9X	73.3	156.3
M9X	74.0	156.3
BM3.3/PBM1-K^b	71.9	154.0
BM3.3/VSV8-K^b	69.7	154.5

[1]Mean positions, $\bar{r}_c$, of the contact distributions shown in Figure 2-3 and Figure 2-4 show a noticeably larger change in $\bar{r}_c$ upon mutation from P2CA to P2CA-A5 (C8A vs. C8P). These data are displayed visually in Figure 2-5. The bottom two rows are calculated from crystal structures obtained by Reiser, et al. [19, 100] and compare the $\bar{r}_c$ values of two TCR/allo-PMHC complexes with the same TCR (BM3.3) and MHC (K^b), but different peptides (PBM1 and VSV8). Comparing these crystal structures to C9X and M9X, peptide mutation is shown to impact $\bar{r}_c$ more than CDR3$_\alpha$ mutation, in qualitative agreement with our simulation results. Errors reported for the simulations are s.e.m.

Bibliography

[1] Kranz, D. M.; Sherman, D. H.; Sitkovsky, M. V.; Pasternack, M. S.; Eisen, H. N. *Proceedings of the National Academy of Sciences of the USA* **1984**, *81*, 573.

[2] Unanue, E. R. *Annual Review of Immunology* **1984**, *2*, 395–428.

[3] McConnell, H. M.; Wada, H. G.; Arimilli, S.; Fok, K. S.; Nag, B. *Proceedings of the National Academy of Sciences of the USA* **1995**, *92*, 2750.

[4] Fowler, K. D.; Kuchroo, V. K.; Chakraborty, A. K. *PloS One* **2012**, *7*, e33018.

[5] Abel, S. M.; Roose, J. P.; Groves, J. T.; Weiss, A.; Chakraborty, A. K. *Journal of Physical Chemistry B* **2012**, *116*, 3630–40.

[6] Govern, C. C.; Yang, M.; Chakraborty, A. K. *Physical Review Letters* **2012**, *108*, 058102.

[7] Won, J.-H.; Goldberger, O.; Shen-Orr, S. S.; Davis, M. M.; Olshen, R. A. *Proceedings of the National Academy of Sciences of the USA* **2012**, *109*, 2848–53.

[8] Dahirel, V.; Shekhar, K.; Pereyra, F.; Miura, T.; Artyomov, M.; Talsania, S.; Allen, T. M.; Altfeld, M.; Carrington, M.; Irvine, D. J.; Walker, B. D.; Chakraborty, A. K. *Proceedings of the National Academy of Sciences of the USA* **2011**, *108*, 11530–5.

[9] Wolfson, M. Y.; Nam, K.; Chakraborty, A. K. *Journal of Physical Chemistry B* **2011**, *115*, 8317–27.

[10] Rivet, C. A.; Hill, A. S.; Lu, H.; Kemp, M. L. *Molecular and Cellular Proteomics* **2011**, *10*.

[11] Ferguson, A. L.; Panagiotopoulos, A. Z.; Debenedetti, P. G.; Kevrekidis, I. G. *Proceedings of the National Academy of Sciences of the USA* **2010**, *107*, 13597.

[12] Košmrlj, A.; Chakraborty, A. K.; Kardar, M.; Shakhnovich, E. I. *Physical Review Letters* **2009**, *103*, 68103.

[13] Alder, B. J.; Wainwright, T. E. *Journal of Chemical Physics* **1959**, *31*, 459.

[14] Frenkel, D.; Smit, B. *Understanding Molecular Simulation: from Algorithms to Applications*, 2nd ed.; Academic: San Diego, Calif. ; London, 2002.

[15] Brooks, B. R.; Bruccoleri, R. E.; Olafson, B. D.; States, D. J.; Swaminathan, S.; Karplus, M. *Journal of Computational Chemistry* **1983**, *4*, 187–217.

[16] Brooks, C. L.; Karplus, M.; Pettitt, B. M. *Proteins: a theoretical perspective of dynamics, structure, and thermodynamics*; Advances in chemical physics ; v. 71; J. Wiley: New York, 1988.

[17] Karplus, M.; McCammon, J. A. *Nature Structural Biology* **2002**, *9*, 646–652.

[18] Garboczi, D. N.; Ghosh, P.; Utz, U.; Fan, Q. R.; Biddison, W. E.; Wiley, D. C. *Nature* **1996**, *384*, 134–141.

[19] Reiser, J. B.; Darnault, C.; Grégoire, C.; Mosser, T.; Mazza, G.; Kearney, A.; van der Merwe, P. A.; Fontecilla-Camps, J. C.; Housset, D.; Malissen, B. *Nature Immunology* **2003**, *4*, 241–247.

[20] Pai, E. F.; Krengel, U.; Petsko, G. A.; Goody, R. S.; Kabsch, W.; Wittinghofer, A. *EMBO Journal* **1990**, *9*, 2351–9.

[21] Arkin, M. R.; Randal, M.; DeLano, W. L.; Hyde, J.; Luong, T. N.; Oslob, J. D.; Raphael, D. R.; Taylor, L.; Wang, J.; McDowell, R. S.; Wells, J. A.; Braisted, A. C. *Proceedings of the National Academy of Sciences of the USA* **2003**, *100*, 1603–8.

[22] Freddolino, P. L.; Arkhipov, A. S.; Larson, S. B.; McPherson, A.; Schulten, K. *Structure (London, England : 1993)* **2006**, *14*, 437–49.

[23] MacKerell Jr, A. D.; et al. *Journal of Physical Chemistry B* **1998**, *102*, 3586–3616.

[24] Jones, J. E. *Proceedings of the Royal Society A* **1924**, *106*, 463–477.

[25] Sethi, A.; Goldstein, B.; Gnanakaran, S. *PLoS Computational Biology* **2011**, *7*, e1002192.

[26] Louet, M.; Perahia, D.; Martinez, J.; Floquet, N. *Journal of Molecular Biology* **2011**, *411*, 298–312.

[27] Levin, A. M.; Bates, D. L.; Ring, A. M.; Krieg, C.; Lin, J. T.; Su, L.; Moraga, I.; Raeber, M. E.; Bowman, G. R.; Novick, P.; Pande, V. S.; Fathman, C. G.; Boyman, O.; Garcia, K. C. *Nature* **2012**, *484*, 529.

[28] Wan, S. Z.; Coveney, P. V.; Flower, D. R. *Journal of Immunology* **2005**, *175*, 1715–1723.

[29] Zoete, V.; Michielin, O. *Proteins: Structure, Function, and Bioinformatics* **2007**, *67*, 1026–1047.

[30] Michielin, O.; Karplus, M. *Journal of Molecular Biology* **2002**, *324*, 547–569.

[31] Andrienko, N.; Andrienko, G. *Exploratory Analysis of Spatial And Temporal Data: A Systematic Approach*, xvi ed.; Springer: Berlin, 2006.

[32] Kellar, K. L.; Kalwar, R. R.; Dubois, K. A.; Crouse, D.; Chafin, W. D.; Kane, B. E. *Cytometry* **2001**, *45*, 27–36.

[33] Perfetto, S. P.; Chattopadhyay, P. K.; Roederer, M. *Nature Reviews Immunology* **2004**, *4*, 648–55.

[34] Janes, K. A.; Kelly, J. R.; Gaudet, S.; Albeck, J. G.; Sorger, P. K.; Lauffenburger, D. A. *Journal of Computational Biology* **2004**, *11*, 544–561.

[35] Janes, K. A.; Albeck, J. G.; Gaudet, S.; Sorger, P. K.; Lauffenburger, D. A.; Yaffe, M. B. *Science* **2005**, *310*, 1646–1653.

[36] Carrari, F.; Baxter, C.; Usadel, B.; Urbanczyk-Wochniak, E.; Zanor, M.-I.; Nunes-Nesi, A.; Nikiforova, V.; Centero, D.; Ratzka, A.; Pauly, M.; Sweetlove, L. J.; Fernie, A. R. *Plant Physiology* **2006**, *142*, 1380–96.

[37] Kemp, M. L.; Wille, L.; Lewis, C. L.; Nicholson, L. B.; Lauffenburger, D. A. *Journal of Immunology* **2007**, *178*, 4984.

[38] Jolliffe, I. T. *Principal Component Analysis*, 2nd ed.; Springer Series in Statistics; Springer-Verlag: New York, 2002.

[39] Gabriel, K. R. *Biometrika* **1971**, *58*, 453–467.

[40] Draper, N. R.; Smith, H. *Applied Regression Analysis*; Wiley Series in Probability and Statistics, Vol. 1; Wiley, 1998.

[41] Kendall, M. G. *A Course in Multivariate Analysis*, 2nd ed.; Hafner Pub. Co.: New York, 1957.

[42] de Jong, S. *Chemometrics and Intelligent Laboratory Systems* **1993**, *18*, 251–263.

[43] Wold, S.; Sjöström, M.; Eriksson, L. *Chemometrics and Intelligent Laboratory Systems* **2001**, *58*, 109–130.

[44] Martens, H.; Martens, M. *Multivariate analysis of quality: an introduction*; John Wiley & Sons Inc., 2001.

[45] Trygg, J.; Wold, S. *Journal of Chemometrics* **2003**, *17*, 53–64.

[46] Wold, H. *Research Papers in Statistics* **1966**, *630*, 411–444.

[47] Zou, X.; Zhao, J.; Mao, H.; Shi, J.; Yin, X.; Li, Y. *Applied Spectroscopy* **2010**, *64*, 786–94.

[48] Hasegawa, K.; Funatsu, K. *SAR and QSAR in Environmental Research* **2000**, *11*, 189–209.

[49] Xiaobo, Z.; Jiewen, Z.; Povey, M. J. W.; Holmes, M.; Hanpin, M. *Analytica Chimica Acta* **2010**, *667*, 14–32.

[50] Boulesteix, A.-L.; Strimmer, K. *Briefings in Bioinformatics* **2007**, *8*, 32–44.

[51] Hasegawa, K.; Miyashita, Y.; Funatsu, K. *Journal of Chemical Information and Computer Sciences* **1997**, *37*, 306–310.

[52] Kimura, T.; Hasegawa, K.; Funatsu, K. *Journal of Chemical Information and Computer Sciences* **1998**, *38*, 276–282.

[53] Hasegawa, K.; Kimura, T.; Funatsu, K. *Journal of Chemical Information and Computer Sciences* **1999**, *39*, 112–120.

[54] Gao, H.; Lajiness, M. S.; Drie, J. V. *Journal of Molecular Graphics and Modeling* **2002**, *20*, 259–268.

[55] Hasegawa, K.; Kimura, T.; Funatsu, K. *Quantitative Structure-Activity Relationships* **1999**, *18*, 262–272.

[56] Wold, S. *Chemometrics and Intelligent Laboratory Systems* **1998**, *44*, 175–185.

[57] Trygg, J.; Wold, S. *Journal of Chemometrics* **2002**, *16*, 119–128.

[58] Malinowski, E. R. *Factor analysis in chemistry*, 2nd ed.; Wiley, 2002.

[59] Löfstedt, T.; Trygg, J. *Journal of Chemometrics* **2011**, 441–455.

[60] Yu, H.; MacGregor, J. F. *Chemometrics and Intelligent Laboratory Systems* **2004**, *73*, 199–205.

[61] Ergon, R. *Journal of Chemometrics* **2005**, *19*, 1–4.

[62] Kemsley, E. K.; Tapp, H. S. *Journal of Chemometrics* **2009**, *23*, 263–264.

[63] Hogquist, K. A.; Jameson, S. C.; Bevan, M. J. *Current Opinions in Immunology* **1994**, *6*, 273–278.

[64] Jameson, S. C.; Hogquist, K. A.; Bevan, M. J. *Annual Review of Immunology* **1995**, *13*, 93–126.

[65] Werlen, G.; Hausmann, B.; Naeher, D.; Palmer, E. *Science* **2003**, *299*, 1859.

[66] Starr, T. K.; Jameson, S. C.; Hogquist, K. A. *Annual Review of Immunology* **2003**, *21*, 139–176.

[67] von Boehmer, H.; Aifantis, I.; Gounari, F.; Azogui, O.; Haughn, L.; Apostolou, I.; Jaeckel, E.; Grassi, F.; Klein, L. *Immunological Reviews* **2003**, *191*, 62–78.

[68] Hogquist, K. A.; Baldwin, T. A.; Jameson, S. C. *Nature Reviews Immunology* **2005**, *5*, 772–782.

[69] Siggs, O. M.; Makaroff, L. E.; Liston, A. *Current Opinions in Immunology* **2006**, *18*, 175–183.

[70] Košmrlj, A.; Jha, A. K.; Huseby, E. S.; Kardar, M.; Chakraborty, A. K. *Proceedings of the National Academy of Sciences of the USA* **2008**, *105*, 16671–6.

[71] Huseby, E. S.; White, J.; Crawford, F.; Vass, T.; Becker, D.; Pinilla, C.; Marrack, P.; Kappler, J. W. *Cell* **2005**, *122*, 247–260.

[72] Huseby, E. S.; Crawford, F.; White, J.; Marrack, P.; Kappler, J. W. *Nature Immunology* **2006**, *7*, 1191–1199.

[73] Lindahl, K. F.; Wilson, D. B. *Journal of Experimental Medicine* **1977**, *145*, 508–522.

[74] Lynes, M. A.; Flaherty, L.; Michaelson, J.; Collins, J. J.; Rinchik, E. M. *Journal of Immunogenetics* **1984**, *11*, 189–196.

[75] Lechler, R. I.; Lombardi, G.; Batchelor, J. R.; Reinsmoen, N.; Bach, F. H. *Immunology Today* **1990**, *11*, 83–88.

[76] Kaufman, C. L.; Gaines, B. A.; Ildstad, S. T. *Annual Review of Immunology* **1995**, *13*, 339–367.

[77] Joyce, S.; Nathenson, S. G. *Immunological Reviews* **1996**, *154*, 59–103.

[78] Le, N. T.; Chen, B. J.; Chao, N. J. *Cytotherapy* **2005**, *7*, 126–133.

[79] Hauben, E.; Bacchetta, R.; Roncarolo, M. G. *Cytotherapy* **2005**, *7*, 158–165.

[80] Felix, N. J.; Allen, P. M. *Nature Reviews Immunology* **2007**, *7*, 942–953.

[81] Nikolich-Zugich, J. *Nature Immunology* **2007**, *8*, 388–397.

[82] Colf, L. A.; Bankovich, A. J.; Hanick, N. A.; Bowerman, N. A.; Jones, L. L.; Kranz, D. M.; Garcia, K. C. *Cell* **2007**, *129*, 135–146.

[83] DeLano, W. L.; *The PyMOL User's Manual*; Palo Alto, CA, USA; 2002.

[84] Alexander-Miller, M. A.; Burke, K.; Koszinowski, U. H.; Hansen, T. H.; Connolly, J. M. *Journal of Immunology* **1993**, *151*, 1–10.

[85] Basu, D.; Horvath, S.; Matsumoto, I.; Fremont, D. H.; Allen, P. M. *Journal of Immunology* **2000**, *164*, 5788–5796.

[86] Eisen, H. N. *Annual Review of Immunology* **2001**, *19*, 1–21.

[87] Huseby, E. S.; Crawford, F.; White, J.; Kappler, J.; Marrack, P. *Proceedings of the National Academy of Sciences of the USA* **2003**, *100*, 11565–11570.

[88] Panina-Bordignon, P.; Corradin, G.; Roosnek, E.; Sette, A.; Lanzavecchia, A. *Science* **1991**, *252*, 1548–50.

[89] Udaka, K.; Tsomides, T. J.; Eisen, H. N. *Cell* **1992**, *69*, 989–998.

[90] Garcia, K. C.; Adams, E. J. *Cell* **2005**, *122*, 333–336.

[91] Armstrong, K. M.; Piepenbrink, K. H.; Baker, B. M. *Biochemical Journal* **2008**, *415*, 183–196.

[92] Felix, N. J.; Donermeyer, D. L.; Horvath, S.; Walters, J. J.; Gross, M. L.; Suri, A.; Allen, P. M. *Nature Immunology* **2007**, *8*, 388–397.

[93] Hornell, T. M.; Martin, S. M.; Myers, N. B.; Connolly, J. M. *Journal of Immunology* **2001**, *167*, 4207–4214.

[94] Garcia, K. C.; Degano, M.; Pease, L. R.; Huang, M.; Peterson, P. A.; Teyton, L.; Wilson, I. A. *Science* **1998**, *279*, 1166–1172.

[95] Ding, Y. H.; Baker, B. M.; Garboczi, D. N.; Biddison, W. E.; Wiley, D. C. *Immunity* **1999**, *11*, 45–56.

[96] Degano, M.; Garcia, K. C.; Apostolopoulos, V.; Rudolph, M. G.; Teyton, L.; Wilson, I. A. *Immunity* **2000**, *12*, 251–261.

[97] Hahn, M.; Nicholson, M. J.; Pyrdol, J.; Wucherpfennig, K. W. *Nature Immunology* **2005**, *6*, 490–496.

[98] Borbulevych, O. Y.; Piepenbrink, K. H.; Gloor, B. E.; Scott, D. R.; Sommese, R. F.; Cole, D. K.; Sewell, A. K.; Baker, B. M. *Immunity* **2009**, *31*, 885–896.

[99] Burrows, S. R.; et al. *Proceedings of the National Academy of Sciences of the USA* **2010**, *107*, 10608.

[100] Reiser, J. B.; Darnault, C.; Guimezanes, A.; Gregoire, C.; Mosser, T.; Schmitt-Verhulst, A. M.; Fontecilla-Camps, J. C.; Malissen, B.; Housset, D.; Mazza, G. *Nature Immunology* **2000**, *1*, 291–297.

[101] Luz, J. G.; Huang, M.; Garcia, K. C.; Rudolph, M. G.; Apostolopoulos, V.; Teyton, L.; Wilson, I. A. *Journal of Experimental Medicine* **2002**, *195*, 1175.

[102] Reiser, J. B.; Grégoire, C.; Darnault, C.; Mosser, T.; Guimezanes, A.; Schmitt-Verhulst, A. M.; Fontecilla-Camps, J. C.; Mazza, G.; Malissen, B.; Housset, D. *Immunity* **2002**, *16*, 345–354.

[103] Archbold, J. K.; Macdonald, W. A.; Miles, J. J.; Brennan, R. M.; Kjer-Nielsen, L.; McCluskey, J.; Burrows, S. R.; Rossjohn, J. *Journal of Biological Chemistry* **2006**, *281*, 34324.

[104] Jones, L. L.; Colf, L. A.; Bankovich, A. J.; Stone, J. D.; Gao, Y. G.; Chan, C. M.; Huang, R. H.; Garcia, K. C.; Kranz, D. M. *Biochemistry* **2008**, *47*, 12398.

[105] Jones, L. L.; Colf, L. A.; Stone, J. D.; Garcia, K. C.; Kranz, D. M. *Journal of Immunology* **2008**, *181*, 6255.

[106] Gras, S.; Burrows, S. R.; Kjer-Nielsen, L.; Clements, C. S.; Liu, Y. C.; Sullivan, L. C.; Bell, M. J.; Brooks, A. G.; Purcell, A. W.; McCluskey, J.; Rossjohn, J. *Immunity* **2009**, *30*, 193–203.

[107] MacDonald, W. A.; et al. *Immunity* **2009**, *31*, 897–908.

[108] Wu, L. C.; Tuot, D. S.; Lyons, D. S.; Garcia, K. C.; Davis, M. M. *Nature* **2002**, *418*, 552–556.

[109] Eisen, H. N.; Chakraborty, A. K. *Proceedings of the National Academy of Sciences of the USA* **2010**, *107*, 22373.

[110] Brünger, A. T. *Proteins: Structure, Function, and Genetics* **1988**, *4*, 148–156.

[111] Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. *Journal of Chemical Physics* **1983**, *79*, 926–935.

[112] Brooks, B. R.; et al. *Journal of Computational Chemistry* **2009**, *30*, 1545–1614.

[113] Mackerell Jr, A. D.; Feig, M.; Brooks III, C. L. *Journal of Computational Chemistry* **2004**, *25*, 1400–1415.

[114] van Gunsteren, W. F.; Berendsen, H. J. C. *Molecular Physics* **1977**, *34*, 1311–1327.

[115] Steinbach, P. J.; Brooks, B. R. *Journal of Computational Chemistry* **1994**, *15*, 667–683.

[116] Essmann, U.; Perera, L.; Berkowitz, M. L.; Darden, T.; Lee, H.; Pedersen, L. G. *Journal of Chemical Physics* **1995**, *103*, 8577–8593.

[117] Nosé, S. *Journal of Chemical Physics* **1984**, *81*, 511–519.

[118] Hoover, W. G. *Physical Review A* **1985**, *31*, 1695–1697.

[119] Feller, S. E.; Zhang, Y.; Pastor, R. W.; Brooks, B. R. *Journal of Chemical Physics* **1995**, *103*, 4613.

[120] Eisen, H. N.; Sykulev, Y.; Tsomides, T. J. *Advances in Protein Chemistry* **1996**, *49*, 1–56.

[121] Kageyama, S.; Tsomides, T. J.; Fukusen, N.; Papayannopoulos, I. A.; Eisen, H. N.; Sykulev, Y. *Journal of Immunology* **2001**, *166*, 3028.

[122] Sha, W. C.; Nelson, C. A.; Newberry, R. D.; Kranz, D. M.; Russell, J. H.; Loh, D. Y. *Nature* **1988**, *336*, 73–76.

[123] Humphrey, W.; Dalke, A.; Schulten, K. *Journal of Molecular Graphics* **1996**, *14*, 33–38.

[124] Gattinoni, L.; Powell, D. J.; Rosenberg, S. A.; Restifo, N. P. *Nature Reviews Immunology* **2006**, *6*, 383–93.

[125] TreeStar Software; *FlowJo*; 2012. http://www.flowjo.com/.

[126] Naumann, T.; Schiller, H. *Formulae and Methods in Experimental Data Evaluation* **1984**.

[127] Knijnenburg, T. A.; Roda, O.; Wan, Y.; Nolan, G. P.; Aitchison, J. D.; Shmulevich, I. *Molecular Systems Biology* **2011**, *7*, 531.

[128] Tumeh, P. C.; Koya, R. C.; Chodon, T.; Graham, N. A.; Graeber, T. G.; Comin-Anduix, B. n.; Ribas, A. *Journal of Immunotherapy* **2010**, *33*, 759–68.

[129] Comin-Anduix, B. n.; Ribas, A. *Unpublished* **2011**.

[130] Trygg, J. *Journal of Chemometrics* **2002**, *16*, 283–293.

[131] Athanassakis, I.; Iconomidou, B. *Developmental Immunology* **1995**, *4*, 247–255.

[132] Nakane, A.; Numata, A.; Minagawa, T. *Infection and Immunity* **1992**, *60*, 523–528.

[133] Goldberger, O.; Sauer, J.-D.; Davis, M. M. *Unpublished* **2011**.

[134] Jeffers, J. N. R. *Journal of the Royal Statistical Society C* **1967**, *16*, 225 – 236.

[135] Bylesjö, M.; Eriksson, D.; Kusano, M.; Moritz, T.; Trygg, J. *Plant Journal* **2007**, *52*, 1181–91.

[136] Abdi, H. In *Encyclopedia of Social Sciences Research Methods*; Lewis-Beck, M.; Bryman, A.; Futing Liao, T., Eds.; Sage: Thousand Oaks, CA, 2004; Vol. 3, pp 792–795.

[137] Zhang, K.; Kim, S.; Cremasco, V.; Hirbe, A. C.; Collins, L.; Piwnica-Worms, D.; Novack, D. V.; Weilbaecher, K.; Faccio, R. *Cancer Research* **2011**, *71*, 4799–808.

[138] Yang, Z.-Z.; Grote, D. M.; Ziesmer, S. C.; Manske, M. K.; Witzig, T. E.; Novak, A. J.; Ansell, S. M. *Blood* **2011**, *118*, 2809–20.

[139] Neurath, M. F.; Hildner, K.; Becker, C.; Schlaak, J. F.; Barbulescu, K.; Germann, T.; Schmitt, E.; Schirmacher, P.; Haralambous, S.; Pasparakis, M.; Meyer Zum Büschenfelde, K. H.; Kollias, G.; Märker-Hermann, E. *Clinical and Experimental Immunology* **1999**, *115*, 42–55.

[140] Sauer, J.-D.; Pereyre, S.; Archer, K. A.; Burke, T. P.; Hanson, B.; Lauer, P.; Portnoy, D. A. *Proceedings of the National Academy of Sciences of the USA* **2011**, *108*, 12419–24.

[141] Kirkland, K. L.; Sillito, A. M.; Jones, H. E.; West, D. C.; Gerstein, G. L. *Journal of Neurophysiology* **2000**, *84*, 1863–8.

[142] Langermans, I. A.; van Furth, R. *Biotherapy* **1994**, *7*, 169–78.

[143] Pamer, E. G. *Nature Reviews Immunology* **2004**, *4*, 812–23.

[144] Conlan, J. W. *Journal of Medical Microbiology* **1996**, *44*, 295.

[145] Geginat, G.; Grauling-Halama, S. In *Handbook of Listeria Monocytogenes*; CRC Press, 2008; Chapter 14, pp 397–426.

[146] Deluca, L. S.; Gommerman, J. L. *Nature Reviews Immunology* **2012**, *12*, 339–51.

[147] Kitamura, M. *Journal of Immunology* **1997**, *159*, 1404–1411.

[148] Shakoory, B.; Fitzgerald, S. M.; Lee, S. A.; Chi, D. S.; Krishnaswamy, G. *Journal of Interferon and Cytokine Research* **2004**, *24*, 271–281.

[149] Gekara, N. O.; Weiss, S. *Cellular Microbiology* **2007**, *10*, 225–36.

[150] Witte, C. E.; Archer, K. A.; Rae, C. S.; Sauer, J.-D.; Woodward, J. J.; Portnoy, D. A. *Advances in Immunology* **2012**, *113*, 135–56.

[151] The Mathworks; *Matlab*; 2011. http://mathworks.com/.

[152] Wold, S. *Technometrics* **1978**, *20*, 397 – 405.

[153] Srinivasan, J.; Cheatham III, T. E.; Cieplak, P.; Kollman, P. A.; Case, D. A. *Journal of the American Chemical Society* **1998**, *120*, 9401–9409.

[154] Kollman, P. A.; Massova, I.; Reyes, C.; Kuhn, B.; Huo, S.; Chong, L.; Lee, M.; Lee, T.; Duan, Y.; Wang, W.; Donini, O.; Cieplak, P.; Srinivasan, J.; Case, D. A.; Cheatham III, T. E. *Accounts of Chemical Research* **2000**, *33*, 889–897.